

Learning Analytics With Neural Networks: Addressing Open Challenges Through Uncertainty Quantification and Natural Language Processing

ARTHUR THUY

Supervisor: Prof. dr. Dries F. Benoit



A dissertation submitted to Ghent University
in partial fulfilment of the requirements
for the degree of
Doctor of Business Economics

Academic year 2025-2026

Typeset in L^AT_EX.

Copyright © 2025 by Arthur Thuy (arthur.thuy@ugent.be)

All rights are reserved. No part of this publication may be reproduced or transmitted in any form or by any means electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the author.

DOCTORAL ADVISORY COMMITTEE

Prof. dr. Dries Benoit (Supervisor, Ghent University)

Prof. dr. ir. Joni Dambre (Professor, Ghent University)

Prof. dr. ir. Katrien Verbert (Professor, KU Leuven)

EXAMINATION BOARD

Prof. dr. Gert Peersman (Chair, Ghent University)

Prof. dr. Matthias Bogaert (Secretary, Ghent University)

Prof. dr. Dries F. Benoit (Supervisor, Ghent University)

Prof. dr. Seppe vanden Broucke (Professor, Ghent University)

Prof. dr. ir. Joni Dambre (Professor, Ghent University)

Prof. dr. ir. Tinne De Laet (Professor, KU Leuven)

Prof. dr. María Óskarsdóttir (Professor, University of Southampton)

LIST OF PUBLICATIONS AND CONFERENCE PRESENTATIONS BASED ON THIS DOCTORAL RESEARCH

Articles

Thuy, A., & Benoit, D. F. (2024). *Explainability through uncertainty: Trustworthy decision-making with neural networks*. European Journal of Operational Research.

Thuy, A., Loginova, E., & Benoit, D. F. (2024). *Active Learning to Guide Labeling Efforts for Question Difficulty Estimation*. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Second International Tutorial and Workshop on Responsible Knowledge Discovery in Education.

Thuy, A., Loginova, E., & Benoit, D. F. (2025). *Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and Ordered-LogitNN*. International Conference on Artificial Intelligence in Education (AIED), Second Workshop on Automated Evaluation of Learning and Assessment Content.

Working papers

Thuy, A., Benedetto, L., Loginova, E., & Benoit, D. F. *Simulating Student Interactions for Virtual Pretesting with In-Context Learning*. Submitted to the Language Resources and Evaluation Conference (LREC 2026).

Conference presentations

Thuy, A., & Benoit, D. F. (2022). *Leveraging uncertainty estimation in neural networks for trustworthy predictions in decision-making*. European Conference on Operational Research (EURO), Helsinki, Finland.

Thuy, A., Loginova, E., & Benoit, D. F. (2024). *Active Learning to Guide Labeling Efforts for Question Difficulty Estimation*. European Conference on Operational Research (EURO), Copenhagen, Denmark.

Thuy, A., Loginova, E., & Benoit, D. F. (2024). *Active Learning to Guide Labeling Efforts for Question Difficulty Estimation*. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Second International Tutorial and Workshop on Responsible Knowledge Discovery in Education, Vilnius, Lithuania.

Thuy, A., Loginova, E., & Benoit, D. F. (2025). *Ordinal regression for question difficulty estimation with transformers*. The Belgian Operational Research Society (ORBEL), Maastricht, the Netherlands.

Thuy, A., Loginova, E., & Benoit, D. F. (2025). *Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and Ordered-LogitNN*. International Conference on Artificial Intelligence in Education (AIED), Second Workshop on Automated Evaluation of Learning and Assessment Content, Palermo, Italy.

Articles not directly related to this dissertation

Thuy, A., & Benoit, D. F. (2024). *Fast and reliable uncertainty quantification with neural network ensembles for industrial image classification*. Annals of Operations Research.

Conference presentations not directly related to this dissertation

Thuy, A., & Benoit, D. F. (2024). *Cheaper neural network ensembles: a case study in manufacturing*. The Belgian Operational Research Society (ORBEL), Antwerp, Belgium.

Acknowledgements

First and foremost, I would like to thank my PhD supervisor, Dries. Thank you for giving me the opportunity to start this PhD, for the enjoyable collaboration over the years, the countless rounds of constructive feedback, and your positive, open-minded attitude during our discussions. You taught me how to conduct research critically and rigorously, identifying shortcomings and building towards a solution in a structured way. Beyond that, I learned essential soft skills from you: how to communicate my research effectively and reflect on my own work—finding the balance between defending my choices and acknowledging that alternative approaches might have been better. These skills have not only shaped me as a researcher but have also made me more mature and confident as a person, something I will take with me for years to come.

I would also like to thank the members of the Examination Board for their insightful comments and for challenging me with critical questions. Although the atmosphere during the internal defense was tense at times, I very much respect how passionate you all are in discussing topics in your area of expertise. What doesn't kill you makes you stronger, and, with some rework, leads to a better PhD thesis.

A special thank you goes to Ekaterina for her guidance throughout the past years. It all started with my master's thesis, where you introduced me to Learning Analytics and Uncertainty Quantification, topics that have remained central throughout my PhD. I truly appreciate the many feedback meetings you managed to fit in during your lunch breaks while balancing a full-time job and other commitments. In addition, I would like to thank Luca for his support in the final study of this dissertation. Working on this project taught me a lot about generative large language models, knowledge that proved invaluable in making the transition to industry.

I am also grateful to my colleagues of the Data Analytics and Consumer Behaviour research clusters for the many enjoyable moments during the lunchtime chats, afterworks at De Geus, and fries at Marcel. I am fortunate to say that a lot of my colleagues have become close friends. In particular, I want to thank my office mates since the very first semester of the PhD, Shimanto and Lukas, the two people who were friends before becoming colleagues. It feels a bit strange to break up the Golden Triangle, but I'll be on the front row of the Faculty Boardroom when your moment arrives.

Dankjewel aan mijn ouders voor jullie onvoorwaardelijke steun, zowel tijdens mijn studiejaren als tijdens het doctoraat. Ik heb jaren lang mogen genieten van het

comfort van *Hotel Mama & Papa*. Elke zondagavond werd ik naar het kot gebracht met een voorraad eten in de hand, maand na maand werd het kot betaald, de was en de plas gedaan, en zoveel meer. Dankzij die zorgeloze periode kon ik mij volledig op mijn studies concentreren. Zonder jullie hulp waren de goede resultaten tijdens de opleiding Handelsingenieur niet mogelijk geweest, en daarmee wellicht ook de kans om aan een doctoraat te beginnen, laat staan het behalen van een FWO-beurs. Ook tijdens het doctoraat waren er meerdere momenten waarop ik zonder jullie steun moeilijk de deadlines had gehaald. Dankzij jullie sta ik hier vandaag; ik hoop dat ik jullie trots heb gemaakt.

Tot slot wil ik mijn vriendin Marieke bedanken voor haar steun tijdens de voorbije jaren. Het doctoraat kwam zelden of nooit ter sprake, en dat was eigenlijk precies wat ik nodig had. Zo kon ik het werk tijdens het weekend helemaal loslaten, om maandag opnieuw met een fris hoofd te beginnen—alsof elk weekend een kleine vakantie was. Nu mijn doctoraat afgelopen is neem ik afscheid van “het schooltje” en ben ik volledig klaar om samen met jou een nieuw hoofdstuk te beginnen. Ik kijk er ontzettend naar uit om de volgende jaren met jou door te brengen!

Ik schrijf deze bedanking met een dubbel gevoel. Enerzijds ben ik ontzettend dankbaar voor iedereen die mij de afgelopen jaren heeft gesteund; anderzijds ben ik bijzonder blij dat het doctoraat tot een einde is gekomen. Al vroeg tijdens het traject had ik door dat ik mij niet thuis voel in de academische wereld. Toch heb ik er nooit aan getwijfeld om het doctoraat af te werken, en ben ik tot mijn laatste werkdag aan de UGent blijven pushen om een sterk resultaat neer te zetten. Laten we klinken op de mooie momenten van de voorbije jaren, en op alles wat nog komen zal. Cheers!

Arthur Thuy

Table of Contents

List of Figures	xiii
List of Tables	xv
Nederlandstalige Samenvatting	xvii
Summary	xix
1 Introduction	1
1.1 Learning Analytics	1
1.1.1 Student Performance Prediction	2
1.1.2 Question Difficulty Estimation	3
1.2 Challenges in Learning Analytics	3
1.3 Fundamental Concepts and Terminology	5
1.3.1 Uncertainty Quantification	5
1.3.2 Natural Language Processing with Large Language Models	7
1.4 Content	9
References	12
2 Explainability through Uncertainty: Trustworthy Decision-Making with Neural Networks	17
2.1 Introduction	18
2.2 Related Work	19
2.2.1 Explainable Artificial Intelligence in Operations Research	19
2.2.2 Uncertainty and Neural Networks in Operations Research	20
2.2.3 Uncertainty as Explainable Artificial Intelligence	21
2.3 Methodology	21
2.3.1 Uncertainty as Explainable Artificial Intelligence	21
2.3.1.1 Data and Model Uncertainty	22
2.3.1.2 Theoretical Properties	23
2.3.1.3 Specification	24
2.3.2 Classification with Rejection	24
2.3.3 Workflow with a Human-in-the-loop	25
2.4 Uncertainty in Neural Networks	25
2.4.1 Data and Model Uncertainty	26
2.4.2 Monte Carlo Dropout	26

2.4.3	Deep Ensembles	27
2.4.4	Uncertainty Decomposition	28
2.5	Case Study: Student Performance Prediction	29
2.5.1	Problem Setting	29
2.5.2	Data	30
2.5.3	Experimental Setup	30
2.6	Results	31
2.6.1	Total Uncertainty	31
2.6.2	Accuracy	34
2.6.3	Non-rejected Accuracy	34
2.6.4	Classification Quality	35
2.6.5	Rejection Quality	36
2.7	Discussion	36
2.8	Conclusion	37
	References	45
3	Active Learning to Guide Labeling Efforts for Question Difficulty Estimation	53
3.1	Introduction	54
3.2	Related Work	55
3.3	Methodology	56
3.3.1	Active Learning	56
3.3.2	Monte Carlo Dropout Uncertainty	57
3.3.3	PowerVariance Acquisition	58
3.4	Experiments	60
3.4.1	Data	60
3.4.2	Model Architecture	60
3.4.3	Active Learning Setup	61
3.5	Results and Discussion	62
3.5.1	Predictive Performance	63
3.5.2	Acquisition Behavior	64
3.6	Conclusion	67
	References	70
4	Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and OrderedLogitNN	77
4.1	Introduction	78
4.2	Related Work	79
4.2.1	Output Types	81
4.2.2	Metrics	81
4.3	Balanced Discrete Ranked Probability Score	82
4.4	Ordered Logit for NNs	85
4.5	Existing Approaches for Ordinal Regression with NNs	86
4.5.1	Discretized Regression	87
4.5.2	Classification	87

4.5.3	Ordinal: OR-NN	87
4.5.4	Ordinal: CORAL and CORN	89
4.6	Experiments	90
4.6.1	Data	90
4.6.2	Model Architecture	90
4.7	Results and Discussion	91
4.7.1	Balanced DRPS	91
4.7.2	Confusion Matrix	92
4.8	Conclusion	94
	References	96
5	Leveraging Misconceptions with In-Context Learning to Simulate Students with Role-Playing LLMs	101
5.1	Introduction	102
5.2	Related Work	103
5.3	Methodology	105
5.3.1	Collecting Skills and Misconceptions	105
5.3.2	Student Behavior Replication	106
5.3.3	Student Behavior Roleplay	107
5.3.4	In-context Example Selection Strategies	107
5.3.5	IRT Estimation	108
5.4	Experimental Setup	109
5.4.1	Dataset	109
5.4.2	Evaluation Metrics	109
5.4.3	Configurations	110
5.5	Results and Analysis	111
5.5.1	Student Behavior Replication	111
5.5.1.1	Predictive performance	111
5.5.1.2	LLM Answer Correctness	112
5.5.2	Student Behavior Roleplay	115
5.6	Conclusions and Future Work	116
	References	123
	Appendix 5.A Prompts	126
	5.A.1 Collecting skills and misconceptions	126
	5.A.2 Student Behavior Replication & Roleplay	127
	Appendix 5.B Additional results	127
	Appendix 5.C Implementation Details	129
6	Conclusion	131
6.1	Overview of Contributions	131
6.2	Critical Reflection	133
6.2.1	Limited Generalization	134
6.2.2	Data Scarcity	135
6.2.3	Misaligned Evaluation Metrics	136
6.2.4	Other Challenges	136

6.3	Implications for Stakeholders	138
6.3.1	Teachers and Instructors	138
6.3.2	Students	139
6.3.3	School Boards and Administrators	139
6.3.4	Educational Software Vendors	140
6.3.5	Policy-Makers and Regulators	140
6.4	Limitations and Future Research	141
	References	143

List of Figures

2.1	Overview explainable artificial intelligence	21
2.2	General uncertainty framework	22
2.3	Two types of uncertainty	23
2.4	Performance measures for classification with rejection	24
2.5	Overview of uncertainty estimation methods	27
2.6	Experimental setup	30
2.7	Standard: uncertainty distributions	32
2.8	MC Dropout: uncertainty distributions	32
2.9	Deep Ensembles: uncertainty distributions	33
2.10	Classification with rejection	35
3.1	Active learning workflow with pool-based sampling	57
3.2	Top- K acquisition toy example	58
3.3	Discrete RMSE as a function of the labeled dataset size	63
3.4	Active gain over Uniform acquisition as a function of the labeled dataset size	64
3.5	Distribution of difficulty levels in the labeled set	65
3.6	Discrete RMSE as a function of the dataset size, per individual difficulty level	66
4.1	Example DRPS calculation with predicted probabilities	84
4.2	Example DRPS calculation with a predicted label	84
4.3	Ordered logit example for three ordinal categories.	86
4.4	Example of a rank inconsistent and rank consistent prediction . . .	88
4.5	Confusion matrices on the ARC dataset.	93
5.1	Overview of the methodology.	104
5.2	LLMs' response correctness over all contextual configurations. . .	113
5.3	LLMs' response correctness for contextual configurations in the llama and qwen3 model families.	114
5.4	LLMs' response correctness for varying hyperparameters of all contextual configurations.	114
5.5	LLMs' validation balanced accuracy for contextual configurations in the llama and qwen3 model families.	129

5.6 LLMs' validation balanced accuracy for varying hyperparameters of all contextual configurations.	130
---	-----

List of Tables

1.1	Learning analytics challenges addressed in this dissertation.	11
1.2	Methodologies used in this dissertation.	11
2.1	Examples of uncertainty decomposition	29
2.2	Accuracy (%)	33
3.1	Hyperparameter settings	62
4.1	Related work on discrete-leveled QDE.	80
4.2	Results on RACE++ (3 levels).	91
4.3	Results on ARC (7 levels).	92
5.1	Best-performing configurations on Student Replication.	111
5.2	Contextual baseline results on Student Replication.	112
5.3	Results on Student Roleplay.	115
5.4	Prompt template for collecting skills and misconceptions.	118
5.5	Prompt template for collecting skills and misconceptions.	127
5.6	Prompt template for student and teacher personas.	128
5.7	Example prompt with variables filled in.	129

Nederlandstalige Samenvatting

Digitale leerplatformen zijn een integraal onderdeel geworden van het moderne onderwijs en genereren uitgebreide data over hoe studenten omgaan met leermateriaal. *Learning analytics* benut deze data om inzichten te verkrijgen in de voortgang en het gedrag van studenten, met als uiteindelijk doel het verbeteren van leerresultaten. Belangrijke toepassingen zijn het voorspellen van studieresultaten voor vroegtijdige waarschuwingssystemen en het aanbevelen van passende oefeningen, waarbij nauwkeurige inschatting van de moeilijkheidsgraad van vragen (question difficulty estimation; QDE) essentieel is.

Ondanks vooruitgang in machinaal leren (machine learning; ML) wordt de toepassing ervan in het onderwijs beperkt door drie aanhoudende uitdagingen: (1) het waarborgen van robuuste generalisatie, (2) omgaan met dataschaarste, en (3) het afstemmen van evaluatiemaatstaven op onderwijsbehoeften. Dit proefschrift pakt deze uitdagingen aan via vier studies, waarbij gebruik wordt gemaakt van methoden uit *onzekerheidsberekening* en *natuurlijke taalverwerking* (natural language processing; NLP) om de robuustheid en relevantie van ML-oplossingen in het onderwijs te vergroten. Drie van deze studies zijn gepubliceerd in peer-reviewed wetenschappelijke tijdschriften of internationale conferenties, terwijl de vierde studie is ingediend voor een conferentie.

De eerste studie onderzoekt hoe onzekerheidsberekening de betrouwbaarheid van voorspellingsmodellen voor studieprestaties kan verbeteren. De resultaten tonen aan dat het identificeren en doorverwijzen van gevallen met hoge onzekerheid naar menselijke experts de veiligheid en geloofwaardigheid van ML-ondersteunde besluitvorming in onderwijscontexten versterkt. De tweede studie richt zich op de hoge data-eisen van QDE en stelt een *active learning*-benadering voor met een nieuwe acquisitiefunctie, PowerVariance. Deze methode behaalt goede prestaties met slechts een fractie van de gelabelde data, wat zorgt voor een meer efficiënte en toegankelijke inzet van QDE-systemen. De derde studie bekijkt QDE vanuit het perspectief van ordinale regressie, waarmee de natuurlijke ordening van moeilijkheidsniveaus wordt benadrukt. Door de introductie van de OrderedLogitNN-architectuur en de balanced Discrete Ranked Probability Score (DRPS) worden zowel de modellering als de evaluatie van QDE-systemen beter onderbouwd, wat leidt tot effectievere en relevantere methodes. De vierde studie onderzoekt of het toevoegen van context over eerdere studentenantwoorden generatieve grote taalmodellen (LLM's) helpt om studentengedrag beter te simuleren. De resultaten tonen aan dat hoewel contextuele informatie het model verrijkt, huidige LLM's nog steeds moeite hebben om authentieke interacties betrouwbaar te repliceren of

te genereren, waardoor hun toepasbaarheid in onderwijs beperkt blijft.

Over alle belanghebbenden heen komt een centrale boodschap naar voor: ML in het onderwijs moet er niet op gericht zijn menselijke expertise te vervangen, maar deze te versterken door middel van onzekerheidsberekening, transparantie en efficiëntie. Leraren blijven de uiteindelijke beslissingsnemers, studenten krijgen duidelijker inzicht in hun leerproces, schoolbesturen krijgen instrumenten om middelen verantwoord op te schalen, softwareleveranciers kunnen betrouwbare en kosteneffectieve systemen ontwikkelen, en beleidsmakers kunnen de resultaten gebruiken in veldstudies voor op bewijs gebaseerde principes voor regulering.

Tot slot draagt dit proefschrift bij aan het overbruggen van de kloof tussen technische vooruitgang in ML en de praktische vereisten van onderwijssystemen. De bevindingen pleiten voor een benadering van learning analytics die niet alleen innovatief is, maar ook verantwoord, efficiënt en in lijn met menselijke waarden.

Summary

Digital learning platforms have become an integral part of modern education, producing extensive data on how students engage with learning content. *Learning analytics* leverages these data traces to gain insights into student progress and behavior, with the ultimate goal of improving learning outcomes. Key applications include *predicting student performance* for early-warning systems and recommending suitable exercises, which relies heavily on accurate *question difficulty estimation* (QDE).

Despite advances in machine learning (ML), its adoption in educational contexts is constrained by three persistent challenges: (1) ensuring robust generalization, (2) addressing data scarcity, and (3) aligning evaluation metrics with educational needs. This dissertation tackles these challenges through a series of four studies, drawing on methods from uncertainty quantification (UQ) and natural language processing (NLP) to improve the robustness and relevance of ML solutions in education. Three of these studies have already been published in peer-reviewed scientific journals or international conferences, while the fourth study is submitted to a conference.

The first study investigates how UQ can improve the reliability of student performance prediction models. It shows that identifying and deferring high-uncertainty cases to human experts enhances the safety and leads to more appropriate trust in ML-assisted decision-making for real-world educational contexts. The second study addresses the high data demands of QDE by proposing an active learning approach with a novel acquisition function, PowerVariance. The method achieves near-fully supervised performance with only a fraction of labeled data, supporting more resource-efficient and accessible deployment of QDE systems. The third study frames QDE as an ordinal regression problem, acknowledging the natural ordering of difficulty levels. By introducing the OrderedLogitNN architecture and the balanced Discrete Ranked Probability Score (DRPS), it enables more accurate modeling and establishes principled evaluation practices, resulting in assessment tools that are both effective and relevant. The fourth study examines whether contextualizing generative large language models (LLMs) with students' prior responses can improve their ability to simulate learner behavior. Findings indicate that while in-context information enriches the setup, current LLMs still struggle to reliably replicate and generate authentic student interactions, limiting their suitability for high-stakes assessment scenarios.

Across all stakeholder groups, a central theme emerges: ML in education should not aim to replace human expertise, but to enhance it through uncertainty-

awareness, transparency, and efficiency. Teachers remain the ultimate decision-makers, students gain clearer insights into their learning, school boards gain tools to scale resources responsibly, software vendors can develop trustworthy and cost-effective systems, and policy-makers can use the results for pilot studies aiming to provide evidence-based principles for regulation.

In sum, this dissertation contributes to bridging the gap between technical advances in ML and the practical requirements of education systems. Its findings argue for an approach to learning analytics that is not only innovative, but also responsible, efficient, and aligned with human values.

1

Introduction

In recent years, digital learning platforms—referred to as Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), and Virtual Learning Environments (VLE)—have become increasingly embedded in educational practice (Hernández-Blanco et al., 2019). In parallel, Massive Open Online Courses (MOOCs) have gained popularity, offering open-access, self-paced learning to a global audience via platforms such as Khan Academy, Duolingo, edX, and Coursera.

The widespread adoption of these platforms has rendered the learning process more observable than ever before. These systems play a central role in the information exchange between instructors and students, while also recording detailed logs of user activity. For instance, they capture when and how often students access particular resources, whether their responses to exercises are correct, and how much time they spend reading or watching instructional material. These “digital traces” left by students offer valuable insights into how learners interact with educational content.

1.1 Learning Analytics

Learning analytics is the process of measuring, collecting, analyzing, and reporting data about learners and their learning environments, with the goal of understanding and improving learning (Long & Siemens, 2011). Recent advances in machine learning (ML) have enabled the use of students’

digital footprints to drive substantial transformations in education, fostering innovative teaching and learning strategies (Markauskaite et al., 2022). Through automation, learning analytics facilitates the evaluation of learning activities (e.g., lectures, courses) and assessment materials (e.g., questions, exams), thereby enabling adaptive and personalized educational experiences at a scale unattainable without computational support (Shum & Luckin, 2019). For example, intelligent tutoring systems can provide personalized guidance and automated feedback to learners (Cavalcanti et al., 2021; Jurenka et al., 2024; Maier & Klotz, 2022). Consequently, educational institutions are increasingly adopting learning analytics systems, as ML technologies promise to make education more accessible, scalable, effective, and tailored to individual learners (Macfadyen, 2022).

Among ML approaches, neural networks (NNs) have been widely adopted in learning analytics, operating on both structured and unstructured data (Hernández-Blanco et al., 2019). Structured data typically include tabular information such as student-question interactions (e.g., response times, access frequencies) and demographic attributes (e.g., age, gender). In contrast, unstructured data often consist of textual information, such as question content and students' written responses.

This dissertation focuses on two central applications within learning analytics: *student performance prediction* and *question difficulty estimation*, which we briefly introduce below.

1.1.1 Student Performance Prediction

A major objective in learning analytics is developing an educational early-warning system, identifying at-risk students in order to enable timely and targeted support (Wong & Li, 2020). To this end, a crucial task is *student performance prediction*, i.e., predicting which students are likely to fail based on common academic outcomes such as student dropout, course certification, final course grade, pass/fail status, etc. By analyzing student interaction data, ML models can raise early warnings, triggering timely instructor interventions. The early-warning results can be made available to teachers in dashboards that are easy to work with. Learning analytics dashboards provide a means to assess student information and student performance predictions, track student progress, and decide which intervention to perform to offer targeted support for at-risk students, potentially leading to improved student outcomes (Fernandez Nieto et al., 2022; Williamson & Kizilcec, 2022).

1.1.2 Question Difficulty Estimation

A central task in learning analytics is *question difficulty estimation (QDE)*, also referred to as question calibration, which aims to estimate a numerical or categorical value representing the difficulty of a question.

QDE plays a crucial role in question generation tools. When teachers can quickly obtain a reliable difficulty estimate while creating exam questions, they can adapt the content to achieve the desired challenge level. With the advent of generative large language models (LLMs), educators can now generate extensive sets of exam questions and efficiently filter those of suitable difficulty to refine their selection (Taslimipoor et al., 2024). This capability can significantly reduce the time required for exam design and preparation.

Accurate difficulty estimation is also vital for exercise recommendation systems used in personalized learning environments, where exercises must align with a student’s skill level. According to Vygotsky’s zone of proximal development (Vygotsky, 1978), the range of exercises that optimally support learning is narrow: overly simple tasks risk inducing boredom, while overly complex ones can cause frustration. Inaccurate difficulty estimates may therefore hinder students’ engagement and negatively affect their learning progress.

Similarly, Computer Adaptive Testing (CAT) (Van der Linden & Glas, 2000) depends heavily on QDE. CAT aims to infer a student’s proficiency by dynamically selecting questions whose difficulty matches their estimated skill level. However, research has shown that CAT performance deteriorates significantly when questions are miscalibrated, that is, when difficulty estimates are inaccurate (Chen et al., 2005).

Traditionally, QDE has relied on manual calibration (Attali et al., 2014) or pretesting (Lane et al., 2016), both of which are costly and time-consuming. Recent progress in natural language processing (NLP), particularly through the use of LLMs, has paved the way for automating QDE, enabling scalable and efficient calibration of new questions (Benedetto et al., 2023).

1.2 Challenges in Learning Analytics

In learning analytics, there are multiple challenges that hinder its wider applicability in real-life settings. Mathrani et al. (2021) request more focus on model generalizability, transparency, and ethical issues. Furthermore, the concerns of Alfredo et al. (2024) are in line, encouraging researchers to emphasize human control, safety, reliability, and trustworthiness. Wilson et

al. (2017) draw attention to the implementation of pedagogical and learning theories in learning analytics. From an ML perspective, Karamolegkou et al. (2025) examine the role of NLP in societal domains, among which education. The authors identify five main recurring challenges across the domains that continue to hinder progress: (1) data scarcity and representational bias; (2) misaligned evaluation metrics; (3) safety, privacy, and ethical concerns; (4) limited generalization; and (5) infrastructural constraints.

This work focuses on three overarching themes that hinder the practical adoption of ML in learning analytics: limited generalization, data scarcity, and misaligned evaluation metrics. Although we do not directly address infrastructural challenges, privacy, and ethics, we reflect on the impact of our work on these topics.

Limited generalization. A central challenge in developing ML systems for learning analytics is ensuring their safe deployment and their ability to generalize across contexts, such as different courses, subjects, and student populations. ML models typically assume that the training and deployment data are drawn from the same distribution. However, in educational contexts, this assumption often fails due to temporal shifts between academic years. For example, a model trained to predict student performance in one academic year may perform poorly in the next. NNs, in particular, are known to exhibit overconfidence under such distribution shifts (Guo et al., 2017), posing risks in high-stakes decision-making environments. There is thus a pressing need for ML models that are robust to these shifts and capable of signaling uncertainty. As such, this challenge is very much linked to other topics like safety, reliability, and trustworthiness (Alfredo et al., 2024; Karamolegkou et al., 2025).

Data scarcity. A persistent challenge in learning analytics is the scarcity of high-quality datasets. This problem is particularly acute for marginalized groups (e.g., languages with few speakers or students with cultural differences) and for exam questions, which instructors are reluctant to share due to concerns over content exposure. The data might not be logged well in online learning platforms, is sensitive to privacy concerns or can only be obtained with high annotation costs. Existing datasets are often small and heterogeneous, reflecting the diversity of course structures and assessment formats. For instance, supervised approaches to QDE require thousands of calibrated questions, yet such datasets are rarely available. Given that question difficulty is so context-dependent, it is important to only use the relevant student and question data. The development of methods that work well with smaller datasets is therefore essential to improve the feasibility of ML tools for these marginalized groups or question types, as such broadening the accessibility of learning analytics.

Misaligned evaluation metrics. In learning analytics, it is crucial that the objectives that the ML tools optimize align with the vision of educational professionals. Misaligned evaluation metrics pose a serious risk to the development and deployment of ML in education. When metrics fail to reflect the true structure of a task, they can mislead researchers, hinder progress, and ultimately lead to suboptimal outcomes and harmful decisions. In learning analytics, existing metrics often fail to reflect the ordinal structure of a task. For example, question difficulty labels are inherently ordinal, e.g., “easy”—“medium”—“hard”, yet most studies ignore this ordering. Karamolegkou et al. (2025) focus on capturing human-centered qualities like empathy and cultural sensitivity, and we argue that this sense of ordinality is also a human-centered quality, albeit more of a quantitative skill. There is a clear need for metrics that are better aligned with the specific objectives and structure of ordinal tasks in education, leading to tools with higher adoption rates.

Others. Although we do not directly address other challenges, we discuss the impact of our work on the infrastructural requirements for schools, e.g., the computational expense of learning analytics solutions and the platforms required for stakeholders to interact with the tools. We also discuss the potential issues in privacy and ethics when working with personal data about students and their learning trajectory. Furthermore, Alfredo et al. (2024) emphasize the importance of human control and Karamolegkou et al. (2025) list human-in-the-loop approaches as an opportunity in scaling expert strategies while preserving teacher agency. In this work, we incorporate several human-in-the-loop settings, approaching it more like a method that provides a possible solution as opposed to a challenge.

To address the three main challenges, the studies in this dissertation explore novel techniques from the fields of NLP or uncertainty quantification (UQ), or a combination of both. The following section provides a brief overview of these methods.

1.3 Fundamental Concepts and Terminology

1.3.1 Uncertainty Quantification

NNs are increasingly deployed in real-world applications, including safety-critical domains, making it essential to assess the trustworthiness of their predictions (Kraus et al., 2020). A key challenge in these settings is the prevalence of distribution shifts—situations where the data encountered during deployment diverges from the distribution seen during training. These shifts may occur suddenly, gradually, or seasonally (Huyen, 2022).

In education, for example, academic years naturally introduce gradual shifts due to changes in course content, student cohorts, and institutional policies.

Ovadia et al. (2019) demonstrate that when faced with distribution shifts, NNs often make highly confident but incorrect predictions. In such cases, the model fails to recognize that it does not know, a phenomenon known as overconfidence. Ideally, the model should express high uncertainty under these conditions, effectively raising a flag for human intervention. The field of UQ provides principled methods for modeling and interpreting such uncertainty, playing a critical role in the safe deployment of NNs in production.

Uncertainty arises from two fundamentally distinct sources: aleatoric and epistemic uncertainty (Der Kiureghian & Ditlevsen, 2009).

Aleatoric uncertainty, or data uncertainty, stems from the inherent noise or randomness in the data-generating process. It is irreducible—even with unlimited data—because it reflects variability that cannot be explained by the model. A classic example is coin flipping: no amount of additional observations can eliminate the 50/50 randomness in the outcomes of a fair coin (Hüllermeier & Waegeman, 2021). As such, the best a model can do is to assign probabilities to possible outcomes (e.g., “heads” or “tails”) without ever producing a definite prediction.

Epistemic uncertainty, or model uncertainty, reflects the model’s lack of knowledge about the best model parameters, typically due to insufficient or unrepresentative training data. Unlike aleatoric uncertainty, epistemic uncertainty is reducible through the acquisition of new information. For instance, a student can be unsure whether the correct sentence is “their test scores are poor” or “there test scores are poor”. The uncertainty is not due to randomness but due to lack of knowledge—a web search could easily resolve it. In other words, epistemic uncertainty refers to the reducible part of the uncertainty, whereas aleatoric uncertainty refers to the irreducible part.

In standard NN architectures, these two sources of uncertainty are not handled equally (Hüllermeier & Waegeman, 2021). Traditional NN classifiers capture only aleatoric uncertainty through the softmax layer, and standard regressors offer no uncertainty estimates at all. To capture epistemic uncertainty, one must resort to approximate Bayesian inference methods (such as Monte Carlo dropout (Gal & Ghahramani, 2016)), ensembling techniques (Lakshminarayanan et al., 2017), evidential deep learning (Park et al., 2023), or density- and distance-based methods (Charpentier et al., 2022; Van Amersfoort et al., 2020). UQ is foundational to human-in-the-loop learning strategies such as classification with rejection and active learning:

Classification with rejection allows a model to abstain from making predictions when uncertainty is high, deferring those cases to human experts. This is particularly valuable in settings where incorrect predictions are costly or harmful. By rejecting only the most uncertain instances, the system balances automation with safety, enhancing the trustworthiness of ML-assisted decision-making.

Active learning is a data-efficient modeling paradigm in which a model is trained iteratively on a small labeled dataset, selectively querying human annotators only for the most informative unlabeled examples (Settles, 2009). This informativeness is closely tied to epistemic uncertainty, which is naturally higher in regions of the input space that are underrepresented in the training data. As a result, active learning prioritizes label acquisition where it is most likely to learn new information.

In sum, UQ provides the theoretical and practical tools necessary for deploying NNs in high-stakes, real-world scenarios such as education, where decisions based on ML models can have significant consequences for students.

1.3.2 Natural Language Processing with Large Language Models

In education, there is an abundance of textual data, including exam questions, multiple-choice options, and open-ended student responses. NLP, a field at the intersection of linguistics and ML, is particularly valuable as it allows computers to understand and generate human language. NLP enables scalable analysis of these texts, uncovering patterns in student understanding, misconceptions, and learning progress. By leveraging such insights, educators can support applications like QDE and automated feedback. In recent years, NLP has been revolutionized with the advent of LLMs—NNs trained on vast corpora of text that can perform a wide array of language tasks, often with minimal task-specific supervision.

This progress in universal language understanding and generation has been driven by two key developments: the introduction of Transformer architectures (Vaswani et al., 2017) and advances in large-scale self-supervised learning. These innovations enable the development of foundation models, which are pre-trained on massive unlabeled datasets and later adapted to a range of downstream tasks. Well-known examples include the BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) model families. The key insight behind this paradigm is the separation of general language modeling from task-specific learning: rather than training separate models for education from scratch, we pre-train a general-purpose model and

fine-tune or prompt it for specific educational applications.

NLP objectives generally fall into two categories:

- **Sequence encoding**, where the model produces a vector representation of a given input sequence. Encoder-only architectures (e.g., BERT) are commonly used in this setting, and the learned representations can be fed into downstream models for tasks such as classifying whether a student’s answer is correct, predicting question difficulty, or classifying misconceptions in written responses.
- **Sequence generation**, where the model generates new sequences based on a given context. Decoder-only architectures (e.g., GPT) excel in this setting and are well-suited for tasks such as generating automated feedback, generating question distractors, or simulating student responses in role-playing scenarios.

Once pre-trained, LLMs can be adapted to specific tasks using two primary approaches: fine-tuning and prompting.

Fine-tuning of pre-trained encoding models. Fine-tuning involves adapting a pre-trained encoder (typically trained with masked language modeling) to a supervised task. During pre-training, a softmax layer is added to the encoder to predict masked tokens based on both left and right context, resulting in a bidirectional language model. For downstream tasks, this softmax layer is replaced with a task-specific output head (e.g., a classifier or regressor), and the model is further trained on labeled data. This process adjusts the parameters of the pre-trained model to better suit the task at hand.

Because pre-training is computationally expensive, fine-tuning offers a practical alternative: it requires only a modest amount of labeled data and limited computational resources. This efficiency has made fine-tuning the standard method for adapting encoder-based LLMs to a wide range of applications in NLP, including learning analytics.

Prompting of pre-trained generation models. Decoder-only Transformer architectures, such as those used in the GPT family, are typically trained using causal language modeling (Radford et al., 2018). These models learn to predict the next token in a sequence based on its left-hand context, ignoring future tokens during training. As a result, they excel at generating coherent and contextually appropriate text.

These models can then be adapted to new tasks with prompting—reframing an NLP problem as a text completion task. For instance, given a student’s open-ended answer to a math problem, one can prompt an LLM with an instruction such as: “Classify the following response as correct, partially correct, or incorrect:”, followed by the student’s answer. The

model then completes the prompt with the appropriate label. This approach enables zero-shot learning, allowing the model to perform tasks it was never explicitly trained on. An extension of this idea is few-shot learning, commonly implemented via in-context learning (ICL). In this setting, the prompt includes several demonstrations—examples of inputs paired with their correct outputs—before presenting the model with a new input to process. These demonstrations serve to teach the model how to perform the task without any additional training.

1.4 Content

Three of the studies in this dissertation have been peer-reviewed and published in scientific journals or international conferences; the fourth is currently under revision.

Chapter 2, “Explainability through Uncertainty: Trustworthy Decision-Making with Neural Networks”, examines how UQ can enhance the reliability of student performance prediction models. NNs have been widely used for this task; however, in practice, models are typically trained on data from one cohort of students and then applied to predict outcomes for subsequent cohorts or different courses. This setup naturally introduces a distribution shift, as both the student population and course content evolve over time. NNs, however, are known to produce overconfident yet inaccurate predictions under such shifts, leading to unnoticed degradation in performance, a challenge that has received limited attention in prior research. To address this issue, the chapter proposes a general framework that uses UQ to determine when model predictions can be trusted. The framework further integrates UQ estimates into a classification with rejection scheme, enabling the system to automatically decide whether to accept a prediction or defer to a human expert. Beyond reliability, UQ offers insight into model confidence and behavior, thus serving as a form of explainable AI (XAI). Overall, this framework fosters more appropriate trust in machine learning systems and supports more reliable downstream decision-making in educational contexts.

The following chapters focus on QDE, the task of assessing the difficulty level of exam questions. Chapter 3, “Active Learning to Guide Labeling Efforts for Question Difficulty Estimation”, addresses the substantial data requirements of current NLP-based QDE approaches. While fine-tuning encoder-based Transformer models yields state-of-the-art performance, these methods typically require large labeled datasets comprising thousands of calibrated questions. For educators seeking to implement QDE tools, collecting such extensive labeled data is often impractical or

infeasible. In contrast, existing unsupervised methods eliminate the need for labeled data but rely on alternative evaluation paradigms that are less convenient in practice. This chapter bridges the gap between supervised and unsupervised methods by introducing an active learning approach to QDE, a human-in-the-loop strategy that begins with a small labeled dataset and iteratively selects the most informative questions for annotation. To guide this selection process, the chapter draws on UQ techniques introduced in Chapter 2 and proposes the novel PowerVariance acquisition function, which reduces redundancy when selecting question batches for labeling. Experimental results show that active learning achieves performance close to fully supervised models while requiring only a fraction of the labeled data, thereby enabling more efficient use of educational resources and broadening access to practical QDE tools for teachers and learning platforms.

Chapter 4, “Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and OrderedLogitNN”, underscores the importance of recognizing the ordinal nature of QDE. Although discrete difficulty levels are inherently ordered (e.g., “easy”–“medium”–“hard” or CEFR levels A1, A2, B1, etc.), prior research has largely neglected this property in both model design and evaluation metrics. Most existing approaches treat QDE as a standard classification or discretized regression problem, overlooking methods developed specifically for ordinal regression. Moreover, widely used evaluation metrics are often model-dependent, fail to capture ordinality, and inadequately address class imbalance. As a result, current ML models tend to optimize for metrics that misalign with how educators intuitively reason about question difficulty, leading to less meaningful predictions and limiting practical adoption. The diversity of evaluation metrics across studies further complicates cross-paper comparability, impeding progress in the field. To overcome these limitations, the chapter introduces both a new evaluation metric and a modeling framework for discrete-level QDE. First, it benchmarks classification, regression, and ordinal regression models using the proposed Balanced DRPS, a metric that simultaneously accounts for ordinality and class imbalance. Second, it presents OrderedLogitNN, which reinterprets the well-established ordered logit model from econometrics as a NN architecture for ordinal regression. Experimental results show that OrderedLogitNN matches baseline models on simpler tasks and outperforms them on more complex ones, while Balanced DRPS provides a more principled and interpretable evaluation framework. Together, these contributions strengthen the alignment between ML-based QDE tools and educators’ reasoning about difficulty, paving the way for broader and more meaningful adoption in practice.

Table 1.1: Learning analytics challenges addressed in this dissertation.

Challenge	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Limited generalization	✓			
Data scarcity		✓		✓
Misaligned evaluation metrics			✓	

Table 1.2: Methodologies used in this dissertation.

Method	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Uncertainty Quantification	✓	✓		
Natural Language Processing		✓	✓	✓

Chapter 5, “Leveraging Misconceptions with In-Context Learning to Simulate Students with Role-Playing LLMs”, investigates whether providing contextual information improves LLMs’ ability to simulate student behavior. In traditional QDE, question difficulties are determined through pretesting, where student responses are collected and analyzed using psychometric models, a process that is both time-consuming and costly, and risks prematurely exposing exam content. To address these drawbacks, recent research has explored virtual pretesting, in which LLMs generate simulated student responses as a low-cost alternative. However, existing approaches have achieved only limited success. This chapter introduces an ICL framework that enriches simulation by incorporating prior question-answer records, giving LLMs more personalized information about the student they are role-playing. Empirical findings reveal that contextual information does not consistently enhance simulation quality, and that optimal configurations fail to generalize across different LLMs, posing challenges for real-world deployment. Consequently, current LLM-based role-playing techniques remain inadequate for high-stakes educational assessments. Overall, the findings suggest that while LLMs can mimic certain aspects of learner behavior, their ability to replicate authentic student performance remains limited, underscoring the need for behaviorally grounded evaluation frameworks in future research.

Tables 1.1 and 1.2 provide a structured overview of how the individual chapters in this dissertation contribute to addressing the key challenges in applying ML within educational contexts. Table 1.1 maps each chapter to the specific challenges it addresses. Table 1.2 summarizes the primary

methodological approaches employed across the dissertation. UQ is applied in Chapters 2 and 3; NLP forms the methodological foundation in Chapters 4 and 5, and also contributes to Chapter 3, which lies at the intersection of UQ and NLP.

References

- Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and ai in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215.
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series*, 2014(2), 1–8. <https://doi.org/10.1002/ets2.12042>
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1–37. <https://doi.org/10.1145/3556538>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2, 100027.
- Charpentier, B., Borchert, O., Zügner, D., Geisler, S., & Günnemann, S. (2022). Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. <https://openreview.net/forum?id=tV3N0DWMxCg>
- Chen, C.-M., Lee, H.-M., & Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, 44(3), 237–255.
- Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2), 105–112. <https://doi.org/10.1016/j.strusafe.2008.06.020>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Fernandez Nieto, G. M., Kitto, K., Buckingham Shum, S., & Martinez-Maldonado, R. (2022). Beyond the learning analytics dashboard: Alternative ways to communicate student data insights combining visualisation, narrative and storytelling. *LAK22: 12th international learning analytics and knowledge conference*, 219–229.
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, 1050–1059.
- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330.
- Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. *Complexity*, 2019(1), 1306039.
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3), 457–506. <https://doi.org/10.1007/s10994-021-05946-3>
- Huyen, C. (2022). *Designing machine learning systems*. O'Reilly Media, Inc.
- Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., et al. (2024). Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.
- Karamolegkou, A., Borah, A., Cho, E., Choudhury, S. R., Galletti, M., Ghosh, R., Gupta, P., Ignat, O., Kargupta, P., Kotonya, N., et al. (2025). Nlp for social good: A survey of challenges, opportunities, and responsible deployment. *arXiv preprint arXiv:2505.22327*. <https://doi.org/10.48550/arXiv.2505.22327>

- Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications [Featured Cluster: Business Analytics: Defining the field and identifying a research agenda]. *European Journal of Operational Research*, 281(3), 628–641. <https://doi.org/10.1016/j.ejor.2019.09.018>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- Lane, S., Raymond, M. R., Haladyna, T. M., et al. (2016). *Handbook of test development* (Vol. 2). Routledge New York, NY.
- Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review (Online)*.
- Macfadyen, L. P. (2022). Institutional implementation of learning analytics: Current state, challenges, and guiding frameworks. *The handbook of learning analytics*, 2, 173–186.
- Maier, U., & Klotz, C. (2022). Personalized feedback in digital learning environments: Classification framework and literature review. *Computers and education: Artificial intelligence*, 3, 100080.
- Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., De Laat, M., Shum, S. B., Gašević, D., et al. (2022). Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with ai? *Computers and Education: Artificial Intelligence*, 3, 100056.
- Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, 2, 100060.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- Park, Y., Choi, W., Kim, S., Han, D.-J., & Moon, J. (2023). Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. *The Eleventh International Conference on Learning Representations*.

- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- Settles, B. (2009). Active learning literature survey.
- Shum, S. B., & Luckin, R. (2019). Learning analytics and ai: Politics, pedagogy and practices. *British journal of educational technology*, 50(6), 2785–2793.
- Taslimipoor, S., Benedetto, L., Felice, M., & Buttery, P. (2024). Distractor generation using generative and discriminative capabilities of transformer-based models. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5052–5063.
- Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. *International conference on machine learning*, 9690–9700.
- Van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Springer. <https://doi.org/10.1007/0-306-47531-6>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.
- Williamson, K., & Kizilcec, R. (2022). A review of learning analytics dashboard research in higher education: Implications for justice, equity, diversity, and inclusion. *LAK22: 12th international learning analytics and knowledge conference*, 260–270.
- Wilson, A., Watson, C., Thompson, T. L., Drew, V., & Doyle, S. (2017). Learning analytics: Challenges and limitations. *Teaching in Higher Education*, 22(8), 991–1007.
- Wong, B. T.-m., & Li, K. C. (2020). A review of learning analytics intervention in higher education (2011–2018). *Journal of Computers in Education*, 7(1), 7–28.

2

Explainability through Uncertainty: Trustworthy Decision-Making with Neural Networks

This chapter is based on a study published in the European Journal of Operational Research (Thuy & Benoit, 2024). The versions are identical except for minor language adjustments.

Abstract

Uncertainty is a key feature of any machine learning model and is particularly important in neural networks, which tend to be overconfident. This overconfidence is worrying under distribution shifts, where the model performance silently degrades as the data distribution diverges from the training data distribution. Uncertainty estimation offers a solution to overconfident models, communicating when the output should (not) be trusted. Although methods for uncertainty estimation have been developed, they have not been explicitly linked to the field of explainable artificial intelligence (XAI). Furthermore, literature in operations research ignores the actionability component of uncertainty estimation and does not consider distribution shifts. This work proposes a general uncertainty framework, with contributions being threefold: (i) uncertainty estimation in ML models is

positioned as an XAI technique, giving local and model-specific explanations; (ii) classification with rejection is used to reduce misclassifications by bringing a human expert in the loop for uncertain observations; (iii) the framework is applied to a case study on neural networks in educational data mining subject to distribution shifts. Uncertainty as XAI improves the model’s trustworthiness in downstream decision-making tasks, giving rise to more actionable and robust machine learning systems in operations research.

2.1 Introduction

A representation of uncertainty is desirable and is a key feature of any machine learning (ML) model. Uncertainty is particularly important in neural networks (NNs), which tend to be overconfident in their predictions (Guo et al., 2017). That is, a NN classifier often predicts an incorrect label, despite giving a high predicted probability.

This flaw is especially troubling in situations of distribution shift, where the data distribution during deployment diverges from the training data distribution (Murphy, 2022). Although the model performs well when first deployed, its performance degrades over time as the distribution shift increases without warning the decision maker. Distribution shifts happen all the time, either suddenly, gradually, or seasonally (Huyen, 2022). For example, a demand prediction model is affected by a sudden change in the pricing policy of a competitor or when a new competitor enters the market.

The field of uncertainty estimation provides a solution to overconfident models by capturing the uncertainty in both the data and the model. As such, it communicates when a model’s output should (not) be trusted (Ovadia et al., 2019). Building trust is also the cornerstone of the field of explainable artificial intelligence (XAI), which aims to explain the output of black-box models. XAI techniques are commonly used in operations research (OR) to facilitate the human-computer interaction and thereby support decision-making systems (Cabitza et al., 2023).

The related work on uncertainty estimation and XAI has three shortcomings: (i) uncertainty estimation is not explicitly formulated as an XAI technique following the local/global and model-specific/agnostic specification and there is no theoretical motivation on how uncertainty contributes to explainability; (ii) the available work in OR merely monitors the NN uncertainty estimates without acting upon it; (iii) there is a lack of OR applications that examine the influence of distribution shifts on NN uncertainty, as literature only employs benchmark datasets like MNIST.

A general uncertainty framework is proposed, with contributions being

threefold:

1. The framework first positions uncertainty estimation in ML models as an *XAI technique*, giving local and model-specific explanations. To support this, theoretical properties are discussed, arguing that uncertainty estimation fosters appropriate levels of *trust*, and increased *actionability* and *robustness*.
2. The framework then uses *classification with rejection* (Mena et al., 2021) to reduce misclassifications by bringing a human expert in the loop for uncertain observations.
3. The framework is applied to a case study on *neural networks* in educational data mining, with distribution shifts occurring naturally when deploying the model to production.

The remainder of the chapter is organized as follows. Section 2.2 gives an overview of related work and identifies shortcomings. Section 2.3 presents the general uncertainty framework and positions uncertainty estimation as an XAI technique. Section 2.4 discusses how uncertainty is quantified specifically in NN classifiers. In Section 2.5, the case study in educational data mining with NN uncertainty is presented; Section 2.6 gives the results. Finally, Section 2.7 provides a discussion and Section 2.8 gives a conclusion.

2.2 Related Work

This section discusses related work on XAI, uncertainty estimation, and NNs in the field of OR. Furthermore, extant literature on uncertainty estimation as XAI is discussed. Thereby, three main shortcomings in related work are identified.

2.2.1 Explainable Artificial Intelligence in Operations Research

ML models are widely used in OR to solve complex problems (Choi et al., 2018). However, extant literature often focuses on predictive performance which comes at the expense of model explainability. This lack of explainability leads to decision makers' distrust and unwillingness to adopt analytics in decision support systems (Shin, 2021).

The field of XAI refers to techniques that try to explain how a black-box ML model produces its outcomes. Although still limited, XAI techniques

are increasingly adopted in OR applications, e.g., in credit risk (Bastos & Matos, 2022; Sachan et al., 2020), marketing risk (De Caigny et al., 2018; Van Nguyen et al., 2020), supply chain management (Garvey et al., 2015), healthcare (Piri et al., 2017), and jurisprudence (Delen et al., 2021). As such, XAI bridges the gap to organizational decision-makers by providing understanding into a model's predictions and generating actionable insights.

XAI techniques can be organized based on two main criteria (Adadi & Berrada, 2018). It can be global, i.e., characterize the whole dataset (e.g., partial dependence plot), or local, i.e., explain individual predictions (e.g., counterfactual explanations). It can be model-specific, i.e., capable of explaining only a restricted class of models (e.g., random forest variable importance), or model-agnostic, i.e., applicable to any model (e.g., SHAP).

2.2.2 Uncertainty and Neural Networks in Operations Research

Neural networks are rapidly emerging in operations research (OR), with applications such as credit scoring, demand prediction, and outlier detection (Gunnarsson et al., 2021; Kraus et al., 2020; Van Belle et al., 2021; Verboven et al., 2021). Kraus et al. (2020) point to three key challenges that limit the relevance of deep learning in OR: (i) extensive hyperparameter tuning is required, (ii) lack of uncertainty estimation, and (iii) lack of accountability and explainability.

Uncertainty estimation for NNs has been investigated in different domains of OR: predictive maintenance (Kraus & Feuerriegel, 2019), recommender systems (Nahta et al., 2021), finance (Ghahtarani, 2021), stress-level prediction (Oh et al., 2021), transportation (Feng et al., 2022; Zhang & Mahadevan, 2020), predictive process monitoring (Weytjens & De Weerd, 2022), and educational data mining (Yu et al., 2021). In the available work, however, uncertainty estimates are merely monitored as an additional metric, not used in combination with a human expert such as in *classification with rejection* (i.e., shortcoming 1). Ignoring the actionability of this human-machine combination leaves a large part of the added value on the table.

Moreover, there is a lack of literature on the impact of distribution shifts on NN uncertainty estimates with applications in OR (i.e., shortcoming 2). That is, uncertainty estimates are always evaluated on benchmark datasets, e.g., MNIST and not-MNIST, or using artificial distortions, e.g., Gaussian blur (Ovadia et al., 2019).

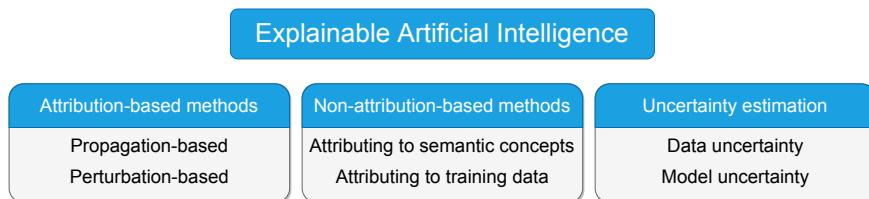


Figure 2.1: **Overview explainable artificial intelligence.** *Uncertainty estimation is a third general type of XAI technique. Figure adapted from Bai et al. (2021).*

2.2.3 Uncertainty as Explainable Artificial Intelligence

Bai et al. (2021) are the first to list uncertainty estimation as a third distinct category in XAI, next to attribution-based (e.g., SHAP) and non-attribution-based (e.g., counterfactual explanations) methods (Figure 2.1). However, Bai et al. (2021) do not specify uncertainty estimation in terms of the two main XAI criteria and do not provide a theoretical motivation (i.e., shortcoming 3).

This work addresses the three shortcomings by proposing a general uncertainty framework, positioning *uncertainty estimation in ML models as XAI* and using *classification with rejection*. Furthermore, a case study on NNs with distribution shifts demonstrates the value of the framework for OR applications.

2.3 Methodology

Figure 2.2 outlines the proposed general uncertainty framework. The framework consists of two stages: (i) uncertainty estimation as XAI and (ii) classification with rejection. The goal of uncertainty estimation is to quantify the data and model uncertainty in predictions made by an ML model. The classification with rejection system then uses the estimates to assist in deciding which predictions should be rejected or retained, based on three key metrics.

2.3.1 Uncertainty as Explainable Artificial Intelligence

Uncertainty as XAI is available for multiple ML models, each having distinct techniques. That is, the case study quantifies data and model uncertainty in NNs, but it can also be computed in e.g., Gaussian Processes (Hüllermeier & Waegeman, 2021) or Random Forests (Shaker & Hüllermeier, 2020) using other existing techniques. Furthermore, one can even

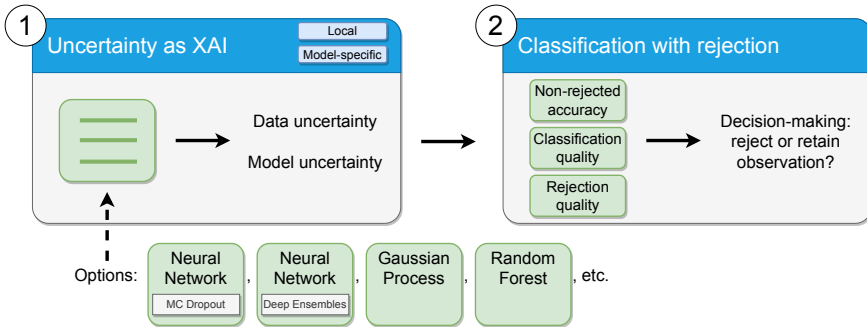


Figure 2.2: **General uncertainty framework.** The framework consists of two stages: (i) uncertainty estimation as XAI and (ii) classification with rejection. It can be applied to multiple ML models, each having one or more specific uncertainty techniques.

use different uncertainty estimation techniques for some ML models, e.g., Monte Carlo Dropout and Deep Ensembles for NNs (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017).

2.3.1.1 Data and Model Uncertainty

Each prediction has two uncertainty values, as uncertainty can arise from two fundamentally different sources: data uncertainty and model uncertainty (Der Kiureghian & Ditlevsen, 2009). *Data uncertainty*, also known as aleatoric uncertainty, refers to the notion of randomness and is related to the data-measurement process. This uncertainty is irreducible even if more data is collected. *Model uncertainty*, also known as epistemic uncertainty, accounts for uncertainty in the model parameters, i.e., uncertainty about which model generated the collected data. In contrast to data uncertainty, collecting more data can reduce model uncertainty. Both types of uncertainty can then be summed to compute the *total uncertainty* in a prediction.

Consider a binary classification task with two input features (Figure 2.3), where the crosses represent positive training examples and the circles represent negative training examples. At test time, predictions are made for both observations *A* and *B*. The model uncertainty is high in sparsely populated regions with few training examples. Therefore, observation *B* has high model uncertainty and could be classified as either positive or negative. In contrast, observation *A* lies in a region where the two class distributions are overlapping, i.e., the data uncertainty is high. Although collecting more training data around observation *B* will reduce the model uncertainty, more training data around observation *A* will not reduce the

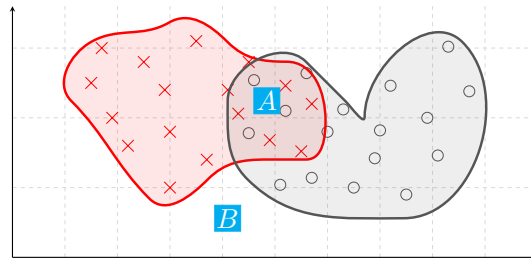


Figure 2.3: **Two types of uncertainty.** Observation A has high data uncertainty; B has high model uncertainty. Figure adapted from Hüllermeier and Waegeman (2021).

data uncertainty.

2.3.1.2 Theoretical Properties

Doshi-Velez and Kim (2017) devise six desirable properties for XAI techniques: trust, actionability, fairness, privacy, robustness, and causality. We apply uncertainty estimation to this list and argue that it satisfies three properties:

- **Trust:** decision makers should feel comfortable relinquishing control to the ML model. As model uncertainty enables saying “I do not know,” a human expert can step in. This awareness gives decision makers more confidence to rely on the model’s predictions in other situations when it says “I do know.”
- **Actionability:** ML models should provide information assisting users to accomplish a task. Uncertainty estimates are key in classification with rejection, where uncertain observations are passed on to a human expert.
- **Robustness:** ML models should reach certain levels of performance in the face of input variation. Under increased distribution shift, uncertainty estimates grow accordingly, enabling an improvement in accuracy by rejecting the most uncertain observations.

To demonstrate the validity of the theoretical properties, they are evaluated in light of the case study results (see Section 2.7). Uncertainty is complementary to other XAI techniques, which can be used to satisfy the remaining three properties.

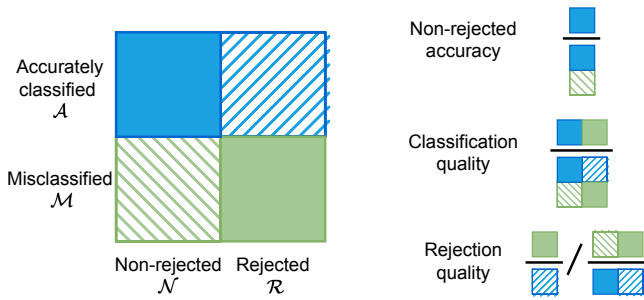


Figure 2.4: **Performance measures for classification with rejection.** Three performance measures are proposed by Condessa et al. (2017) to find the optimal rejection point. Figure adapted from Mena et al. (2021).

2.3.1.3 Specification

Uncertainty estimation is a local and model-specific XAI method. It is *local* because each observation receives an uncertainty estimate, both for data and model uncertainty. Furthermore, it is *model-specific* because techniques for decomposition into data and model uncertainty are different across ML models, although they exist for multiple ML models.

2.3.2 Classification with Rejection

For predictions with high uncertainty, the observations can be passed on to a human expert for a label. The goal of a *classification with rejection* system (Barandas et al., 2022; Mena et al., 2021) is to help decide when to stop rejecting the most uncertain observations. The system takes as input the per-observation uncertainty values and outputs three metrics to assist the decision maker in finding the optimal rejection threshold for the task at hand. It is useful for applications where making an error can be more costly than asking a human expert for help. For example, in fraud detection, an employee can verify a transaction manually if the prediction is uncertain.

Observations are classified along two criteria: (i) accurately classified \mathcal{A} and misclassified \mathcal{M} ; (ii) rejected \mathcal{R} and non-rejected \mathcal{N} . Condessa et al. (2017) propose three rejection metrics (Figure 2.4; higher is better):

- Non-rejected accuracy (NRA): ability to classify non-rejected samples accurately.

$$NRA = \frac{|\mathcal{A} \cap \mathcal{N}|}{|\mathcal{N}|} \quad (2.1)$$

- Classification quality (CQ): ability to retain correctly classified samples and to reject misclassified samples, i.e., correct decision-making.

$$CQ = \frac{|\mathcal{A} \cap \mathcal{N}| + |\mathcal{M} \cap \mathcal{R}|}{|\mathcal{N}| + |\mathcal{R}|} \quad (2.2)$$

- Rejection quality (RQ): ability to concentrate misclassified samples in the set of rejected samples.

$$RQ = \frac{|\mathcal{M} \cap \mathcal{R}|}{|\mathcal{A} \cap \mathcal{R}|} / \frac{|\mathcal{M}|}{|\mathcal{A}|} \quad (2.3)$$

The NRA and CQ are bounded in the interval $[0, 1]$, unlike the RQ which has a minimum value of zero and an unbounded maximum. The three metrics are evaluated as a function of the percentage of rejections varying from 0% to 100% (Yong & Brintrup, 2022). If observations have identical uncertainty values (e.g., exactly 0.0 or 1.0), observations are rejected randomly until the desired percentage is achieved.

2.3.3 Workflow with a Human-in-the-loop

The suggested way of working for the human expert is as follows. The accuracy score is first evaluated on the entire test set, without rejecting any observations. If the accuracy is not sufficiently high, the rejection process is started and the metrics NRA, CQ, and RQ are evaluated. The most uncertain observations are rejected until the labeling budget for the expert annotator is exhausted, or until the NRA is sufficiently high. Although the NRA is the most important metric, the CQ and RQ provide more information on the internals of the rejection system. For example, a decreasing CQ indicates that more and more correct observations are rejected. At this point, the expert might decide to stop rejecting because the NRA will likely stagnate, which does not justify spending the labeling budget on.

It is important to note that the rejection level depends on the labeling budget available for the expert annotator and the accuracy requirement, associated with the misclassification cost, for the task. As such, there is no universally optimal point of rejection.

2.4 Uncertainty in Neural Networks

In the case study, NNs are used for the ‘‘Uncertainty as XAI’’ building block of the general uncertainty framework. This section discusses how uncertainty can be represented and measured in NNs.

2.4.1 Data and Model Uncertainty

Data uncertainty. In a NN classifier, the output layer contains a softmax or sigmoid function, forming a categorical distribution over the class labels $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta})$. This distribution enables the NN to represent *data uncertainty*.

Modern NNs are usually trained using a maximum likelihood objective. That is, they find a single setting of parameters $\boldsymbol{\theta}^*$ to maximize the probability of the data given the parameters, $\arg \max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta})$. For each test input \mathbf{x}^* , there is only one prediction because the NN generates an identical output for each run. As a result, model uncertainty cannot be captured.

Model uncertainty. NNs are large flexible models capable of representing many functions, corresponding to different parameter settings. Each function fits the training data well, yet generalizes in different ways, a phenomenon known as *underspecification* (Wilson, 2020). Considering all of these different NNs together allows capturing *model uncertainty*. In a probabilistic sense, uncertainty in an unseen input point \mathbf{x}^* is represented by the posterior predictive distribution $p(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{Y})$.

Model uncertainty can be captured in NNs using two approaches: (i) Bayesian NNs and (ii) ensembles. A Bayesian NN aims to estimate a full distribution for $p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{Y})$, unlike maximum likelihood. However, this distribution is intractable and is typically approximated using sampling techniques. The ensembling approach obtains multiple good maximum likelihood settings $\boldsymbol{\theta}^*$.

Both approaches aggregate predictions over a collection of NNs. The following subsections discuss the most popular technique for either approach, (i) Monte Carlo Dropout and (ii) Deep Ensembles. Figure 2.5 provides a visual overview.

2.4.2 Monte Carlo Dropout

In Monte Carlo (MC) Dropout (Gal & Ghahramani, 2016), dropout is not only applied at training time but also at test time. Multiple forward passes are performed, each time randomly dropping units and getting another thinned dropout variant of the NN. The resulting T predictions $\{\hat{\mathbf{y}}_1^*(\mathbf{x}^*), \dots, \hat{\mathbf{y}}_T^*(\mathbf{x}^*)\}$ are aggregated, forming an approximation to the

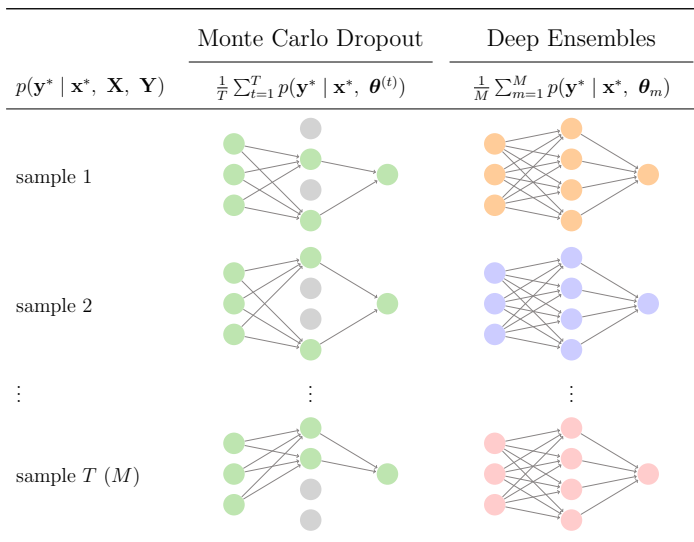


Figure 2.5: **Overview of uncertainty estimation methods.** Forward passes are generated differently depending on the method. In MC dropout, different units are dropped out from a NN; in Deep Ensembles, multiple independent NNs are used, with different parameter initializations and noise in the SGD training process.

true posterior predictive distribution:

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathbf{X}, \mathbf{Y}) d\boldsymbol{\theta} \quad (2.4)$$

$$\approx \frac{1}{T} \sum_{t=1}^T p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}^{(t)}). \quad (2.5)$$

The posterior predictive distribution is obtained through Bayesian model averaging. That is, it averages over an infinite collection of parameter settings, weighted by their posterior probabilities.

2.4.3 Deep Ensembles

Deep Ensembles (Lakshminarayanan et al., 2017) uses an ensemble of M maximum likelihood NNs, with every NN trained on the same dataset and the same input features. The diversity arises through different parameter initializations and noise in the stochastic gradient descent (SGD) training process, inducing different solutions due to the non-convex loss. At test time, each of the M NNs performs one forward pass. The resulting M

predictions $\{\hat{\mathbf{y}}_1^*(\mathbf{x}^*), \dots, \hat{\mathbf{y}}_M^*(\mathbf{x}^*)\}$ are averaged, forming a mixture distribution:

$$p(\mathbf{y}^* | \mathbf{x}^*) \approx \frac{1}{M} \sum_{m=1}^M p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}_m). \quad (2.6)$$

In ensembling, NNs are weighted equally over a finite collection of functions. As such, it is a fundamentally different mindset than Bayesian model averaging.

2.4.4 Uncertainty Decomposition

The posterior predictive distribution holds information about the total uncertainty in a prediction, decomposable in data and model uncertainty using classical information-theoretic measures. However, calculations require the expectation over the posterior distribution, which is intractable. Nonetheless, an approximation can be obtained using samples from the approximate posterior predictive distribution:

$$u_{total}(\mathbf{x}^*) \approx H \left[\frac{1}{T} \sum_{t=1}^T p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}^{(t)}) \right] \quad (2.7)$$

$$u_{data}(\mathbf{x}^*) \approx \frac{1}{T} \sum_{t=1}^T H \left[p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}^{(t)}) \right] \quad (2.8)$$

$$u_{model}(\mathbf{x}^*) = u_{total}(\mathbf{x}^*) - u_{data}(\mathbf{x}^*). \quad (2.9)$$

First, total uncertainty and data uncertainty are calculated; then model uncertainty is obtained as the difference (Depeweg et al., 2018). Total uncertainty is computed by averaging over the different samples and calculating the entropy H . Data uncertainty is computed by calculating the entropy in each sample and averaging the entropies. This boils down to fixing a set of weights $\boldsymbol{\theta}^{(t)}$, i.e., considering a distribution $p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}^{(t)})$, essentially removing the model uncertainty. Model uncertainty is high if the distribution $p(\mathbf{y}^* | \mathbf{x}^*, \boldsymbol{\theta}^{(t)})$ varies greatly for different weights $\boldsymbol{\theta}^{(t)}$. Intuitively, data uncertainty measures uncertainty in the softmax classification on individual samples; model uncertainty measures how much the samples deviate (Barandas et al., 2022; Hüllermeier & Waegeman, 2021).

Table 2.1 contains three examples in the context of binary classification with $T = 4$ samples. The middle and bottom rows both have a total uncertainty of 1.0 although the samples are wildly different. Therefore, the total uncertainty alone is not sufficient to characterize the NN’s predictions; decomposition into data and model uncertainty is necessary.

Table 2.1: Examples of uncertainty decomposition. The middle and bottom row have equal total uncertainty but have wildly different samples. Decomposition in data and model uncertainty reveals the different characteristics.

Samples $p(\mathbf{y}^* \mathbf{x}^*, \boldsymbol{\theta}^{(t)})$	$p(\mathbf{y}^* \mathbf{x}^*)$	$u_{total}(\mathbf{x}^*)$	$u_{data}(\mathbf{x}^*)$	$u_{model}(\mathbf{x}^*)$
$\{(1, ., 0.), (1, ., 0.), (1, ., 0.), (1, ., 0.)\}$	$(1, ., 0.)$	0.	0.	0.
$\{(0.5, 0.5), (0.5, 0.5), (0.5, 0.5), (0.5, 0.5)\}$	$(0.5, 0.5)$	1.	1.	0.
$\{(1, ., 0.), (0, ., 1.), (1, ., 0.), (0, ., 1.)\}$	$(0.5, 0.5)$	1.	0.	1.

2.5 Case Study: Student Performance Prediction

2.5.1 Problem Setting

Student performance prediction is extensively discussed in OR literature (Coussement et al., 2020; Deeva et al., 2022; Delen et al., 2020; Olaya et al., 2020; Phan et al., 2023). Common performance metrics include student dropout, course certification, final course grade, pass/fail, etc. Developing predictive models for student performance forms the basis for an educational early-warning system, where at-risk students are identified on time and assisted with personalized support by course advisors. Therefore, in order to deliver support, a predictive model should provide predictions being both *accurate* and *actionable*.

Whitehill et al. (2017) and Gardner and Brooks (2018) argue that most prior research has poor actionability due to same course–same year evaluation. This training paradigm creates a practical problem because the target labels required by supervised learning algorithms only become available after the final exam, when any support for students is too late. Alternatives that resolve this issue are training on a previous edition of the course, or training on a different course altogether if there is no previous edition available.

The case study applies the uncertainty framework to NNs and investigates uncertainty estimation as an XAI technique in a predictive setup subject to distribution shifts. The results are compared to a standard NN only capable of capturing data uncertainty, but no model uncertainty. The experiment answers the call of Gašević et al. (2016) for research on changing course conditions in student performance prediction, advocating that learning analytics should account for the fluid nature of technology use within a course.

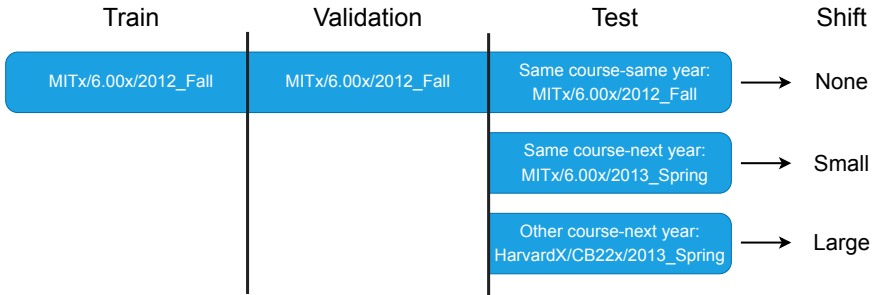


Figure 2.6: **Experimental setup.** A NN is trained on the course “MITx/6.00x/2012_Fall”. Next, predictions are made for three test sets: (i) “MITx/6.00x/2012_Fall” (same course–same year), (ii) “MITx/6.00x/2013_Spring” (same course–next year), (iii) “HarvardX/CB22x/2013_Spring” (other course–next year).

2.5.2 Data

The dataset (MITx & HarvardX, 2014) consists of student-course records of HarvardX and MITx massive open online courses (MOOCs) hosted on the edX platform about a wide range of topics, over two semesters (fall 2012 and spring 2013). The binary target labels denote whether a student scored a grade high enough to earn a certificate; features include processed clickstream activities and student demographics.

2.5.3 Experimental Setup

The experimental setup is detailed in Figure 2.6. First, the NN is trained on the course “MITx/6.00x/2012_Fall”, denoting the MITx course 6.00x “Introduction to Computer Science and Programming” of fall 2012. This course is selected because it has the largest number of observations and is offered in both semesters. Next, predictions are made on three test sets: (i) same course–same year: “MITx/6.00x/2012_Fall”, (ii) same course–next year: “MITx/6.00x/2013_Spring”, and (iii) other course–next year: “HarvardX/CB22x/2013_Spring”. Course CB22x is titled “The Ancient Greek Hero” and is selected as an extreme case because it is a non-STEM course offered by a different university. It is important to note that the input features gathered for HarvardX and MITx courses are identical because they are both hosted on the edX platform.

Predicting on the three test sets represents three different distribution shifts. This ranges from (i) no shift (same course–same year), to (ii) small shift (same course–next year), to (iii) large shift (other course–next year).

The situation of no shift serves as a baseline because it is often used in literature, despite being practically infeasible.

The data include students who accessed at least half of the chapters in the course material, with the training set having a class distribution of 56/44 and 2000 observations. The three test sets have a class distribution of (i) 56/44, (ii) 46/54, and (iii) 72/28, respectively.

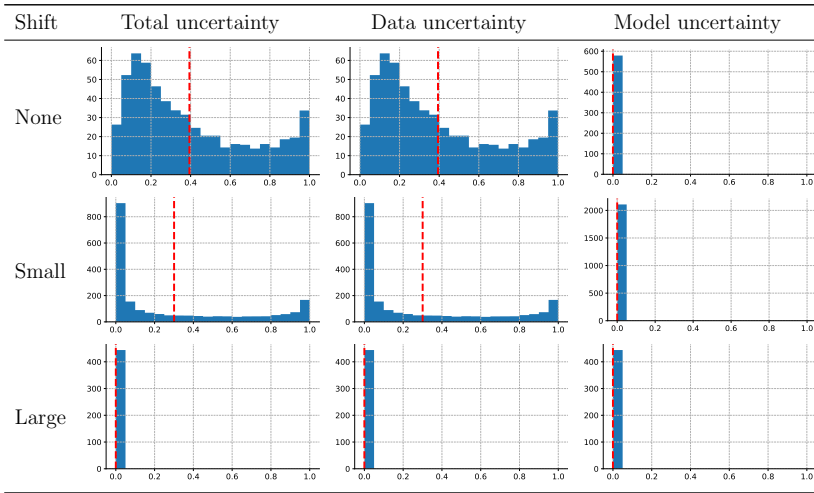
All three methods (standard, MC Dropout, Deep Ensembles) use the same NN configuration as the building block. The NN is a multi-layer perceptron with 2 hidden layers, each containing 64 hidden units, a ReLU activation function, a dropout rate of 0.4 and 0.5, and Glorot uniform weight initialization. The NNs are trained with the Adam optimizer and a binary cross-entropy loss function for 50 epochs with a batch size of 32 and a learning rate of 5×10^{-4} , using early stopping. For MC Dropout, the NN predicts 128 samples per input observation. For Deep Ensembles, 10 NNs are trained, resulting in 10 samples per input observation. Results are averaged over 10 runs with random data splits.

2.6 Results

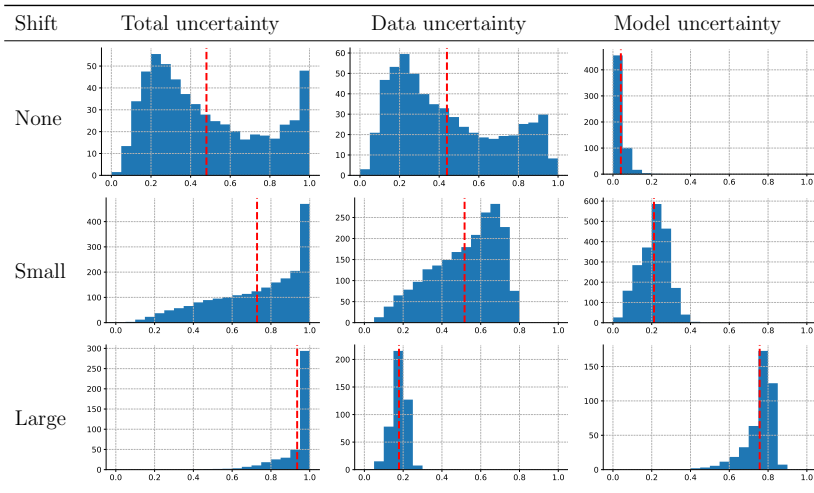
2.6.1 Total Uncertainty

Figures 2.7, 2.8, and 2.9 show the histograms of the uncertainty distributions for all three methods on all three shifts (rows); the total uncertainty is decomposed into data and model uncertainty (columns). The vertical dashed line denotes the mean of the uncertainty values, which can be used to quickly compare centrality across distributions.

With increased distribution shift, data uncertainty consistently decreases for the standard NN, i.e., more mass is located at low uncertainty values and the mean value decreases. For MC Dropout and Deep Ensembles, data uncertainty remains equal when moving to the small shift before decreasing substantially on the large shift. In contrast to data uncertainty, model uncertainty grows rapidly for MC Dropout and Deep Ensembles. The standard NN cannot capture model uncertainty (i.e., value zero for all observations) and only relies on the decreasing data uncertainty to calculate total uncertainty.



*Figure 2.7: **Standard:** uncertainty distributions. The histogram displays absolute frequency and the dashed line denotes the mean value. With increased distribution shift, data uncertainty decreases so total uncertainty decreases as well because the standard NN does not capture model uncertainty.*



*Figure 2.8: **MC Dropout:** uncertainty distributions. The histogram displays absolute frequency and the dashed line denotes the mean value. With increased distribution shift, data uncertainty stagnates or decreases while model uncertainty increases consistently. As a result, MC Dropout has increased total uncertainty.*

In summary, for increased distribution shift, the standard NN has de-

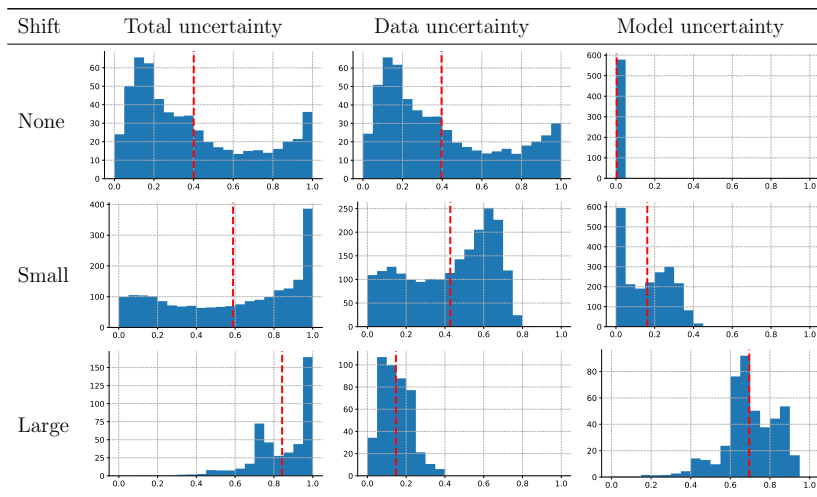


Figure 2.9: Deep Ensembles: uncertainty distributions. The histogram displays absolute frequency and the dashed line denotes the mean value. With increased distribution shift, data uncertainty stagnates or decreases while model uncertainty increases consistently. As a result, Deep Ensembles has increased total uncertainty.

Table 2.2: Accuracy (%). Accuracy degrades with increased distribution shift for all NNs. For a small and large shift, MC Dropout and Deep Ensembles outperform the standard NN. Mean \pm standard error are reported.

Shift	Standard	MC Dropout	Deep Ensembles
None	88.17 ± 0.10	87.81 ± 0.25	88.53 ± 0.11
Small	63.85 ± 0.92	84.43 ± 1.04	84.93 ± 0.52
Large	27.93 ± 0.00	45.05 ± 4.30	31.01 ± 1.57

creased total uncertainty, whereas MC Dropout and Deep Ensembles have rapidly increased total uncertainty. In other words, the standard NN becomes more confident as the inputs stray away from the training data distribution, which is undesirable behavior. This is in contrast to MC Dropout and Deep Ensembles, which indicate that the NN “knows what it does not know.”

2.6.2 Accuracy

Table 2.2 shows the accuracy for all three methods (columns) on all three shifts (rows). It is important to note that MC Dropout and Deep Ensembles average over the different samples to get the final probability vector, capturing model uncertainty.

On increasingly shifted data, the accuracy degrades for all three methods, as expected. For no shift, MC Dropout has a slightly lower accuracy than the standard NN while Deep Ensembles performs slightly better. For a large shift, all NNs perform poorly because they tend to naively predict the minority class, which illustrates how difficult the task is. The results of MC Dropout also have a larger standard error in this situation, indicating that the model results alternate between predicting the minority class and making more sensible predictions. In the situation of a small shift, information contained in different samples (i.e., model uncertainty) has a big impact on accuracy; MC Dropout and Deep Ensembles improve substantially over the standard NN. That is, the standard NN has an accuracy of 63.85%, MC Dropout has 84.43%, and Deep Ensembles has 84.93%.

It is worth noting that although all methods have poor accuracy under a large shift, MC Dropout and Deep Ensembles raise a warning through increased uncertainty whereas the standard NN is confidently wrong. The uncertainty values are then used to reject the most uncertain observation, i.e., classification with rejection.

2.6.3 Non-rejected Accuracy

Figure 2.10 (left column) displays the non-rejected accuracy for all three methods on all three shifts (rows), based on the total uncertainty. Note that the curve at rejection 0.0% corresponds to the method's accuracy without rejection in Table 2.2.

In the situation of no shift (top row), all three methods achieve 100% accuracy. The same holds for a small shift (middle row), despite that the standard NN started at a substantially lower initial accuracy. For a large shift (bottom row), initial accuracies are all poor but only MC Dropout and Deep Ensembles manage to increase the accuracy by rejecting the most uncertain observations. Deep Ensembles performs better with 95% accuracy at rejection rate 0.95, whereas MC Dropout only obtains 80% accuracy. The standard NN, in contrast, has a large amount of observations with uncertainty zero. Since these uncertainty values are identical, observations are rejected randomly, causing the non-rejected accuracy to stagnate at the initial accuracy.

Only MC Dropout and Deep Ensembles have informative uncertainty

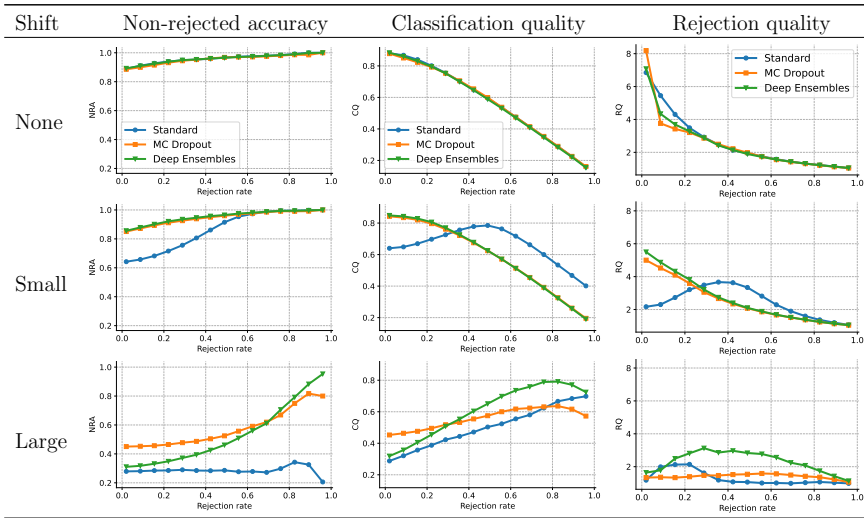


Figure 2.10: **Classification with rejection.** NRA, CQ, and RQ are displayed for increased distribution shifts, for all three models. Under a large distribution shift, the standard NN’s uncertainty estimates are uninformative and rejections are random. This is evidenced by the stagnating NRA, ever-increasing CQ and low RQ.

values so that appropriate observations are rejected, effectively increasing the NRA. The standard NN, on the other hand, fails to increase the NRA under large distribution shifts.

2.6.4 Classification Quality

Figure 2.10 (middle column) shows the classification quality for all three methods on all three shifts (rows), based on the total uncertainty. Classification quality measures the correct decision-making of the classifier-rejector; accurately classified samples should be retained and misclassified samples should be rejected. In other words, the curve shows where the maximum number of correct decisions is made.

In the situation of no shift (top row), all three curves decrease gradually, indicating that the majority of rejected observations are correctly classified. As such, it is optimal to not reject any observations. When a distribution shift is present (middle and bottom row), MC Dropout and Deep Ensembles obtain the point of optimal decision-making at smaller rejection rates than the standard NN. For the small shift (middle row), MC Dropout and Deep Ensembles obtain the maximum CQ at rejection rate 0%; the standard NN needs 50% rejections.

For the large shift (bottom row), the CQ curve of the standard NN keeps on increasing, i.e., it would be best to reject all observations. In contrast, MC Dropout and Deep Ensembles obtain the maximum value at rejection rate 80%. At this rate, Deep Ensembles outperforms MC Dropout with a CQ of 0.80, as compared to MC Dropout’s CQ of 0.63. These findings indicate that the uncertainty estimates of the standard NN are least effective at deciding whether observations should be retained or rejected, and that Deep Ensembles is more effective than MC Dropout.

2.6.5 Rejection Quality

Figure 2.10 (right column) displays the rejection quality for all three methods on all three shifts (rows), based on the total uncertainty. Rejection quality measures the ability to reject misclassified samples. That is, it compares the proportion of misclassified to accurately classified samples on the set of rejected samples with that proportion on the entire data set.

In the situation of no shift (top row), all three curves decrease rapidly, indicating that the proportion of misclassified observations in the rejected set decreases as more observations are rejected, which is undesirable. For the small shift (middle row), MC Dropout and Deep Ensembles obtain the highest RQ at rejection rate 0%. In contrast, the standard NN requires 40% rejections to do the same.

For the large distribution shift (bottom row), Deep Ensembles outperforms the other two methods by achieving much higher RQ values. This finding indicates that the majority of rejected observations are misclassified, effectively improving the NRA. MC Dropout has lower RQ values, with the NRA curve increasing slower. The standard NN falls quickly to RQ value 1.0, indicating that the uncertainty estimates are uninformative and rejections are random. This trend is reflected in the stagnating NRA.

2.7 Discussion

Methods with uncertainty estimation as XAI perform on par or better than the standard NN. If there is no shift, there is no difference between the methods. However, from the point a distribution shift is present, uncertainty estimation leads to optimal decision-making at smaller rejection rates. For small shifts, capturing model uncertainty induces higher initial accuracy and fewer rejections to obtain a specific level of accuracy. For large distribution shifts, it issues a warning about novel observations so the system can reject predictions accordingly, unlike the standard NN.

Uncertainty as XAI fosters appropriate *trust* in an ML system by also

capturing model uncertainty, indicating when an observation lies outside the observed training data. The results show that the model uncertainty values are sensitive to changes in the data distribution, providing an important source of information to the decision maker not available in the standard NN. Furthermore, *robustness* is improved as the total uncertainty grows with increasing distribution shifts, while uncertainty values in the standard NN decrease. Finally, *actionability* is increased by directly using the uncertainty information in a classification with rejection system, raising the NRA even under large distribution shifts.

Continuing on the increased actionability, the decision maker should inspect Figure 2.10 to decide on the appropriate rejection threshold given the specific labeling budget and accuracy requirements. For example in the situation of no shift, it would be sensible to label at most 20% of the observations as the CQ and RQ decrease quickly, resulting in a slowly increasing NRA curve. In contrast, for the large shift, the full labeling budget can be used as the NRA continues to improve.

2.8 Conclusion

This chapter proposes a general uncertainty framework positioning *uncertainty estimation in ML models as an XAI technique*, giving local and model-specific explanations. Furthermore, the framework uses *classification with rejection* to reduce misclassifications by bringing a human expert in the loop for uncertain observations. Finally, the framework is applied to a case study of NNs in educational data mining subject to distribution shifts.

The case study demonstrates that standard NNs only capturing data uncertainty are confidently wrong when confronted with distribution shifts. In contrast, NNs equipped with uncertainty estimation as XAI raise a warning in novel situations through increased model uncertainty, offering a solution to their overconfidence. Deep Ensembles outperform MC Dropout as an XAI technique with higher quality uncertainty estimates, obtaining higher accuracy when rejecting the most uncertain observations. Uncertainty as XAI improves the model's *trustworthiness* in downstream decision-making tasks, giving rise to more *actionable* and *robust* ML systems in OR.

Several directions for future work are possible. The case study only considers knowing the target labels in time due to limitations in the data; studies also satisfying the requirement for input features would help validate the findings. Although this chapter focuses on uncertainty for NN classifiers, uncertainty can also be quantified for NN regression models and other ML models such as Gaussian Processes (Price et al., 2019) and Random Forests (Shaker & Hüllermeier, 2020). Finally, uncertainty as XAI

can be used in active learning, where limited labeled training data is available and the ML system can ask a human expert to label the most uncertain observations (Kadziński & Ciomek, 2021).

Reflections and Clarifications

Trust Calibration and Actionability

In the context of trust in ML systems, the field of Human-Computer Interaction distinguishes two key concepts: *appropriate trust* and *trust calibration* (Mehrotra et al., 2024). Appropriate trust refers to the alignment between users' perceived and the system's actual performance. When appropriate trust is achieved, practitioners rely on correct predictions for downstream decisions and disregard incorrect ones. In contrast, over-trust and under-trust occur when users respectively overestimate or underestimate a system's reliability, leading to misuse (i.e., overreliance on AI tools) or disuse (i.e., neglect or underutilization) (Parasuraman & Riley, 1997). Accordingly, trust calibration denotes the process by which users adjust their expectations of an ML system's reliability and trustworthiness.

We argue that UQ in ML systems is an essential mechanism for fostering appropriate trust among educational professionals. Chapter 2 demonstrates that, in models capturing both aleatoric and epistemic uncertainty, there exists a strong association between high uncertainty estimates and incorrect predictions. This relationship is evidenced by the increasing non-rejected accuracy curves observed under large distribution shift, whereas a standard NN fails to meaningfully filter uncertain predictions. By leveraging UQ, practitioners can use the estimated uncertainty of each prediction to decide whether it should be accepted or rejected, an approach known as classification with rejection.

Within the broader discussion of trust calibration, emphasis is often placed on the adoption and acceptance of AI tools (Afroogh et al., 2024), reflecting the challenges of both misuse and disuse. This aligns directly with the actionable nature of UQ and classification with rejection: given an uncertainty threshold and an estimated uncertainty value, the framework provides clear guidance on whether a prediction should be retained or deferred to a human expert for review. In this sense, UQ delivers actionable, instance-level insights that support more reliable and informed human-AI collaboration.

Desirable Properties of XAI methods

Doshi-Velez and Kim (2017) list six desirable properties of “ML systems”, referring to the combination of an ML model and an XAI method on top of it. The properties are defined as follows:

1. Fairness or unbiasedness: protected groups (explicit or implicit) are not somehow discriminated against.
2. Privacy: the method protects sensitive information in the data.
3. Reliability and robustness: ascertain whether algorithms reach certain levels of performance in the face of parameter or input variation.
4. Causality: the predicted change in output due to a perturbation will occur in the real system.
5. Usable: information that assist users to accomplish a task.
6. Trusted: ML systems have the confidence of human users.

The authors note that “in many cases, formal definitions remain elusive”, meaning that definitions are difficult to achieve. In our work, we rephrase “usability” to “actionability” because this fits better in the scope of the chapter.

Alternative Approaches to UQ for NNs

In Chapter 2, we focus on two approaches that capture both aleatoric and epistemic uncertainty, belonging broadly to either (i) approximate Bayesian or (ii) ensembling techniques. It is worth noting, however, that the distinction between these categories is often subtle. In this work, we describe Monte Carlo (MC) Dropout as an approximate Bayesian method and Deep Ensembles as an ensembling approach. Nonetheless, some studies classify Deep Ensembles as approximate Bayesian (Wilson, 2020), arguing that “Deep ensembles have been mistaken as competing approaches to Bayesian methods, but can be seen as approximate Bayesian marginalization”. Conversely, MC Dropout is sometimes interpreted as an implicit ensembling technique, “dropout may also be interpreted as ensemble model combination where the predictions are averaged over an ensemble of NNs (with parameter sharing)” (Lakshminarayanan et al., 2017).

An alternative approximate Bayesian approach is stochastic variational inference (SVI), which assumes that NN weights follow a Gaussian distribution. While promising in theory, SVI doubles the number of weight

parameters by estimating both the mean (μ) and standard deviation (σ) of the approximating Gaussian and requires a reparameterization trick to enable backpropagation (Kingma & Welling, 2014). Ovadia et al. (2019) benchmarked SVI against MC Dropout and Deep Ensembles, finding that SVI performs well on small datasets (e.g., MNIST, CIFAR) but struggles to scale to larger datasets such as ImageNet or to more complex architectures like long short-term memory (LSTM) NNs.

Beyond these approaches, several other model families have been developed for UQ, including evidential deep learning and distance- or density-based methods. Evidential deep learning adopts a frequentist perspective, introducing specialized loss functions that estimate a second-order distribution over outcomes, from which aleatoric and epistemic uncertainties can be estimated. For instance, in classification tasks, instead of predicting a Bernoulli parameter θ , such models predict the parameters (α, β) of a Beta distribution. Although these methods have gained traction (Park et al., 2023), their ability to represent epistemic uncertainty faithfully has been questioned (Jürgens et al., 2024).

For distance-based methods, deterministic uncertainty quantification (DUQ) (Van Amersfoort et al., 2020) extends the idea of Radial Basis Functions (RBFs) to derive uncertainty estimates from a single deterministic NN. For density-based methods, Natural Posterior Network (Charpentier et al., 2022) fits a single density function in a learned low-dimensional latent space to obtain second-order distributions. While these approaches provide uncertainty estimates via a single forward pass, they remain relatively niche, likely due to their limited probabilistic grounding.

A related but distinct research line focuses on OOD detection (Shafaei et al., 2019). However, fair comparison with the previously mentioned UQ methods is challenging due to differing modeling assumptions. For example, some OOD detection approaches rely on access to a known OOD dataset during training or introduce an additional “none-of-the-above” class.

Other techniques that only capture aleatoric uncertainty, no epistemic uncertainty, are popular in literature. A well-known example is temperature scaling, a post-hoc calibration technique that adjusts softmax outputs using a validation set. Temperature scaling has been shown to produce well-calibrated predictions on i.i.d. test data (Guo et al., 2017), but its performance degrades under distribution shifts, where methods capturing epistemic uncertainty perform better (Ovadia et al., 2019).

Another prominent framework is conformal prediction (CP) (Vovk et al., 2005), a frequentist approach that represents uncertainty through prediction sets that guarantee coverage of the true label with a user-specified

probability. However, Cabezas et al. (2025) observe that applying CP directly to the output of a standard NN fails to capture epistemic uncertainty, often resulting in overconfident predictions in data-sparse regions. Although CP alone does not model epistemic uncertainty, recent research has explored its integration with existing UQ techniques to enhance their reliability. We elaborate on these approaches in the following section.

Alternative Approaches to UQ outside NNs

This study does not aim to advocate for NNs over traditional ML models in student performance prediction. Rather, it seeks to illustrate how such learning analytics tasks inherently involve distribution shifts and how NNs’ overconfidence under these conditions can lead to misleading or even harmful predictions. Our goal is to raise awareness of the risks posed by distribution shifts in ML applications. In this context, it is also valuable to reflect on alternative ML approaches that are specifically designed for tabular data.

A prominent class of such methods not discussed in Chapter 2 are gradient-boosted decision trees (GBDTs). These models can also capture aleatoric and epistemic uncertainty, provided their standard formulations are suitably adapted (Duan et al., 2020). For instance, in CatBoost, a widely used GBDT implementation optimized for categorical input features, aleatoric uncertainty can be modeled in regression settings—something not accounted for in standard regression models (Malinin et al., 2021). Instead of outputting only the mean μ of a homoscedastic Gaussian distribution, CatBoost can predict both the mean μ and the standard deviation σ of a heteroscedastic Gaussian. Since the ground-truth values of σ are not observed, they are learned via a specialized loss function known as learned loss attenuation (implemented as `RMSEWithUncertainty` in the Python package `catboost`). This approach closely parallels the estimation of aleatoric uncertainty in NN regressors (Kendall & Gal, 2017). For classification, CatBoost inherently provides aleatoric uncertainty, requiring no modification.

Epistemic uncertainty can also be derived from an existing GBDT model using a technique known as *virtual ensembles*, which leverages the fact that a GBDT is itself an ensemble of decision trees (Malinin et al., 2021). Because the trees are built sequentially and therefore not fully independent, the virtual ensemble method constructs “truncated” sub-models from the original GBDT and treats these as ensemble members. In CatBoost, this functionality is available through the `virtual_ensembles_predict` function, which serves as a direct al-

ternative to the standard `predict` function.

Reliable and Explainable UQ

Javanmardi et al. (2025) emphasize that, despite their advantages, methods capturing both aleatoric and epistemic uncertainty “lack a formal notion of reliability in the uncertainties they express”. This observation raises important questions about the reliability of existing UQ techniques.

In this regard, CP plays a crucial role, as it provides a principled framework for converting any heuristic measure of uncertainty, regardless of the underlying model, into one with formal statistical guarantees (Angelopoulos & Bates, 2021). The intersection of UQ methods and CP has recently attracted growing attention. For instance, Rossellini et al. (2024) propose incorporating epistemic uncertainty into conformalized quantile regression, while Karimi and Samavi (2024) design a nonconformity score derived from evidential uncertainty estimates. Similarly, Cabezas et al. (2025) compute nonconformity scores using approximate Bayesian models, and Javanmardi et al. (2025) introduce Bernoulli prediction sets (BPS), a set-based approach that applies CP principles to epistemic uncertainty estimates.

Another important direction in UQ concerns the explainability of the uncertainty estimates themselves. Although UQ is positioned here as an XAI technique for communicating when model predictions should or should not be trusted, it does not inherently explain why a given observation exhibits high or low uncertainty. To address this, attribution-based XAI methods for UQ are needed. Research in this area remains scarce: Watson et al. (2023), for example, adapt the Shapley value framework to attribute the contribution of individual features to model uncertainty. We anticipate that greater attention will be devoted to XAI for UQ once the field has addressed its foundational challenges, most notably, ensuring the reliability of uncertainty estimates, as discussed above.

Classification with Rejection

To clarify the classification with rejection workflow, we present pseudocode in Algorithm 1, illustrating how the methodology can be applied during deployment. It is important to note that determining the optimal uncertainty threshold τ during model development requires human judgment in balancing the NRA, CQ, and RQ metrics. As this process involves subjective trade-offs, it cannot be readily formalized as pseudocode.

In this framework, observations are typically rejected based on their total uncertainty, defined as the sum of aleatoric and epistemic components. From a practical standpoint, the specific source of uncertainty, whether

Algorithm 1 Classification with Rejection: Deployment

Require: Trained model \mathcal{M} , uncertainty threshold τ **Require:** Dataset $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ **Ensure:** Final label set $\mathcal{Y} = \{y_1, y_2, \dots, y_N\}$

```

1:  $\mathcal{Y} \leftarrow \emptyset$ 
2: for  $i = 1$  to  $N$  do
3:    $(\hat{y}_i, u_i) \leftarrow \mathcal{M}(\mathbf{x}_i)$   $\triangleright u_i$  represents total uncertainty
4:   if  $u_i > \tau$  then
5:      $y_i \leftarrow \text{HUMANEXPERT}(\mathbf{x}_i)$ 
6:   else
7:      $y_i \leftarrow \hat{y}_i$ 
8:   end if
9:   Append  $y_i$  to  $\mathcal{Y}$ 
10: end for
11: return  $\mathcal{Y}$ 

```

aleatoric or epistemic, is less relevant when the objective is to minimize misclassifications. As long as the most uncertain (and thus most error-prone) predictions are rejected first, the non-rejected accuracy naturally improves.

Sources of Uncertainty

The distinction between aleatoric and epistemic uncertainty is not always clear-cut. Importantly, there are no ground-truth values for either type of uncertainty. Hüllermeier and Waegeman (2021) emphasize that “aleatoric and epistemic uncertainty should not be seen as absolute notions. Instead, they are context-dependent in the sense of depending on the setting $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathcal{P})$. Changing the context will also change the sources of uncertainty: aleatoric may turn into epistemic uncertainty and vice versa”. Here, \mathcal{X} denotes the input space (the set of instances), \mathcal{Y} the output space (the set of outcomes), \mathcal{H} the hypothesis space (the model), and \mathcal{P} the data-generating process. A similar perspective is offered by Der Kiureghian and Ditlevsen (2009), who note that “these concepts only make unambiguous sense if they are defined within the confines of a model of analysis”.

Moreover, aleatoric and epistemic uncertainty are not entirely independent. Consider, for instance, disease detection in X-ray imaging: a blurry image introduces aleatoric uncertainty (noise inherent in the data), while the presence of a rare disease reflects epistemic uncertainty (limited model knowledge). A blurry image of a rare disease combines both sources, mak-

ing it difficult for the model to disentangle uncertainty arising from noise versus ignorance, an interplay that depends on model capacity and data coverage.

Because there are no ground-truth uncertainty values, evaluating the quality of uncertainty estimates remains a challenging problem. A common indirect approach is to assess UQ through classification with rejection. If uncertainty estimates are reliable, incorrect predictions should exhibit high uncertainty, leading to an increase in NRA as the most uncertain observations are progressively rejected.

Neural Network Settings

Hyperparameters for the standard NN architecture were selected through manual tuning, optimizing for validation accuracy. The tuned hyperparameters include the number of hidden layers, number of hidden units, batch size, learning rate, and dropout rate. For MC Dropout, we adopted the same hyperparameter configuration, as its architecture is identical to the standard NN, except that dropout remains active during inference. For Deep Ensembles, each ensemble member was trained using the same hyperparameter settings as the standard NN.

The choice of the number of samples and ensemble members was guided by the benchmarking study of Ovadia et al. (2019). Across their experiments, the authors used different sample sizes for stochastic methods such as MC Dropout, depending on the dataset and NN architecture: for MNIST, 300 and 128 samples (images, 2-layer MLP and LeNet); for CIFAR-10, 128 (images, ResNet-20); for ImageNet, 128 (images, ResNet-50); for Newsgroups, 5 (text, LSTM); and for Criteo, 128 (tabular, 3-layer MLP). Based on these insights, we used 128 samples for MC Dropout, consistent with their configuration for image and tabular data. Their results further indicate that increasing the number of samples enhances performance, though improvements plateau beyond approximately five samples.

For Deep Ensembles, Ovadia et al. (2019) employed 10 ensemble members across all experiments, a setup we also adopt. Similarly to MC Dropout, they observed that performance improves with more ensemble members but exhibits diminishing returns beyond five.

Overall, the benchmarking study of Ovadia et al. (2019), covering diverse datasets and NN architectures, suggests that these UQ methods are robust to variations in model size, sample size, and dataset scale, with no strict requirements on them.

Acknowledgments

This work was supported by the Research Foundation Flanders (FWO) (grant number 1S97022N).

References

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in ai: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), 1–30.
- Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
- Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, *120*, 108102. <https://doi.org/10.1016/j.patcog.2021.108102>
- Barandas, M., Folgado, D., Santos, R., Simão, R., & Gamboa, H. (2022). Uncertainty-based rejection in machine learning: Implications for model development and interpretability. *Electronics*, *11*(3), 396. <https://doi.org/10.3390/electronics11030396>
- Bastos, J. A., & Matos, S. M. (2022). Explainable models of credit losses. *European Journal of Operational Research*, *301*(1), 386–394. <https://doi.org/10.1016/j.ejor.2021.11.009>
- Cabezas, L. M. C., Santos, V. S., Ramos, T., & Izbicki, R. (2025). Epistemic uncertainty in conformal scores: A unified approach. In S. Chiappa & S. Magliacane (Eds.), *Conference on uncertainty in artificial intelligence, rio othon palace, rio de janeiro, brazil, 21-25 july 2025* (pp. 443–470, Vol. 286). PMLR. <https://proceedings.mlr.press/v286/cruz-cabezas25a.html>
- Cabitz, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., & Holzinger, A. (2023). Quod erat demonstrandum?—towards a typology of the concept of explanation for the design of explainable ai. *Expert Systems with Applications*, *213*, 118888. <https://doi.org/10.1016/j.eswa.2022.118888>

- Charpentier, B., Borchert, O., Zügner, D., Geisler, S., & Günnemann, S. (2022). Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. <https://openreview.net/forum?id=tV3N0DWMxCg>
- Choi, T.-M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management*, 27(10), 1868–1883. <https://doi.org/10.1111/poms.12838>
- Condessa, F., Bioucas-Dias, J., & Kovačević, J. (2017). Performance measures for classification systems with rejection. *Pattern Recognition*, 63, 437–450. <https://doi.org/10.1016/j.patcog.2016.10.011>
- Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, 135, 113325. <https://doi.org/10.1016/j.dss.2020.113325>
- De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, 269(2), 760–772. <https://doi.org/10.1016/j.ejor.2018.02.009>
- Deeva, G., De Smedt, J., Saint-Pierre, C., Weber, R., & De Weerd, J. (2022). Predicting student performance using sequence classification with time-based windows. *Expert Systems with Applications*, 209, 118182. <https://doi.org/10.1016/j.eswa.2022.118182>
- Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a bayesian belief network-based dss for predicting and understanding freshmen student attrition [Featured Cluster: Business Analytics: Defining the field and identifying a research agenda]. *European Journal of Operational Research*, 281(3), 575–587. <https://doi.org/10.1016/j.ejor.2019.03.037>
- Delen, D., Zolbanin, H. M., Crosby, D., & Wright, D. (2021). To imprison or not to imprison: An analytics model for drug courts. *Annals of Operations Research*, 303, 101–124. <https://doi.org/10.1007/s10479-021-03984-7>
- Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning

- for efficient and risk-sensitive learning. *International Conference on Machine Learning*, 1184–1193.
- Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2), 105–112. <https://doi.org/10.1016/j.strusafe.2008.06.020>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. <https://doi.org/10.48550/arXiv.1702.08608>
- Duan, T., Anand, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., & Schuler, A. (2020). Ngboost: Natural gradient boosting for probabilistic prediction. *International conference on machine learning*, 2690–2700.
- Feng, R., Ma, A., Jing, Z., Gu, X., Dang, P., & Yao, B. (2022). Understanding the uncertainty of traffic time prediction impacts on parking lot reservation in logistics centers. *Annals of Operations Research*, 1–23. <https://doi.org/10.1007/s10479-022-04734-z>
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, 1050–1059.
- Gardner, J., & Brooks, C. (2018). Student success prediction in moocs. *User Modeling and User-Adapted Interaction*, 28(2), 127–203. <https://doi.org/10.1007/s11257-018-9203-z>
- Garvey, M. D., Carnovale, S., & Yeniyurt, S. (2015). An analytical framework for supply network risk propagation: A bayesian network approach. *European Journal of Operational Research*, 243(2), 618–627. <https://doi.org/10.1016/j.ejor.2014.10.034>
- Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, 28, 68–84. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- Ghahtarani, A. (2021). A new portfolio selection problem in bubble condition under uncertainty: Application of z-number theory and fuzzy neural network. *Expert Systems with Applications*, 177, 114944. <https://doi.org/10.1016/j.eswa.2021.114944>
- Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1), 292–305. <https://doi.org/10.1016/j.ejor.2021.03.006>

- Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330.
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. <https://doi.org/10.1007/s10994-021-05946-3>
- Huyen, C. (2022). *Designing machine learning systems*. O'Reilly Media, Inc.
- Javanmardi, A., Zargarbashi, S. H., Thies, S. M., Waegeman, W., Bojchevski, A., & Hüllermeier, E. (2025). Optimal conformal prediction under epistemic uncertainty. *arXiv preprint arXiv:2505.19033*.
- Jürgens, M., Meinert, N., Bengs, V., Hüllermeier, E., & Waegeman, W. (2024). Is epistemic uncertainty faithfully represented by evidential deep learning methods? *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. <https://openreview.net/forum?id=mxjB0LIgpT>
- Kadziński, M., & Ciomek, K. (2021). Active learning strategies for interactive elicitation of assignment examples for threshold-based multiple criteria sorting. *European Journal of Operational Research*, *293*(2), 658–680. <https://doi.org/10.1016/j.ejor.2020.12.055>
- Karimi, H., & Samavi, R. (2024). Evidential uncertainty sets in deep classifiers using conformal prediction. In S. Vantini, M. Fontana, A. Solari, H. Boström, & L. Carlsson (Eds.), *The 13th symposium on conformal and probabilistic prediction with applications, 9-11 september 2024, politecnico di milano, milano, italy* (pp. 466–489, Vol. 230). PMLR. <https://proceedings.mlr.press/v230/karimi24a.html>
- Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, *30*.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, banff, ab, canada, april 14-16, 2014, conference track proceedings*. <http://arxiv.org/abs/1312.6114>
- Kraus, M., & Feuerriegel, S. (2019). Forecasting remaining useful life: Interpretable deep learning approach via variational bayesian infer-

- ences. *Decision Support Systems*, 125, 113100. <https://doi.org/10.1016/j.dss.2019.113100>
- Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications [Featured Cluster: Business Analytics: Defining the field and identifying a research agenda]. *European Journal of Operational Research*, 281(3), 628–641. <https://doi.org/10.1016/j.ejor.2019.09.018>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- Malinin, A., Prokhorenkova, L., & Ustimenko, A. (2021). Uncertainty in gradient boosting via ensembles. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. <https://openreview.net/forum?id=1Jv6b0Zq3qi>
- Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4), 1–45.
- Mena, J., Pujol, O., & Vitrià, J. (2021). A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9), 1–35. <https://doi.org/10.1145/3477140>
- MITx, [& HarvardX. (2014). Hmxpc13_di_v2_5-14-14.csv. In *Harvardx-mitx person-course academic year 2013 de-identified dataset, version 2.0*. Harvard Dataverse. <https://doi.org/10.7910/DVN/26147/OCLJIV>
- Murphy, K. (2022). Probabilistic machine learning: Advanced topics. MIT Press.
- Nahta, R., Meena, Y. K., Gopalani, D., & Chauhan, G. S. (2021). A hybrid neural variational cf-nade for collaborative filtering using abstraction and generation. *Expert Systems with Applications*, 179, 115047. <https://doi.org/10.1016/j.eswa.2021.115047>
- Oh, B., Hwang, J., Seo, S., Chun, S., & Lee, K.-H. (2021). Inductive gaussian representation of user-specific information for personalized stress-level prediction. *Expert Systems with Applications*, 178, 114912. <https://doi.org/10.1016/j.eswa.2021.114912>
- Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education.

- Decision Support Systems*, 134, 113320. <https://doi.org/10.1016/j.dss.2020.113320>
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), 230–253.
- Park, Y., Choi, W., Kim, S., Han, D.-J., & Moon, J. (2023). Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. *The Eleventh International Conference on Learning Representations*.
- Phan, M., De Caigny, A., & Coussement, K. (2023). A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems*, 113940. <https://doi.org/10.1016/j.dss.2023.113940>
- Piri, S., Delen, D., Liu, T., & Zolbanin, H. M. (2017). A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. *Decision Support Systems*, 101, 12–27. <https://doi.org/10.1016/j.dss.2017.05.012>
- Price, I., Fowkes, J., & Hopman, D. (2019). Gaussian processes for unconstraining demand. *European Journal of Operational Research*, 275(2), 621–634. <https://doi.org/10.1016/j.ejor.2018.11.065>
- Rossellini, R., Barber, R. F., & Willett, R. (2024). Integrating uncertainty awareness into conformalized quantile regression. *International Conference on Artificial Intelligence and Statistics*, 1540–1548.
- Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable ai decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 144, 113100. <https://doi.org/10.1016/j.eswa.2019.113100>
- Shafaei, A., Schmidt, M., & Little, J. J. (2019). A less biased evaluation of out-of-distribution sample detectors. *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 3. <https://bmv2019.org/wp-content/uploads/papers/0333-paper.pdf>
- Shaker, M. H., & Hüllermeier, E. (2020). Aleatoric and epistemic uncertainty with random forests. *Advances in Intelligent Data Analysis*

- XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18, 444–456. https://doi.org/10.1007/978-3-030-44584-3_35
- Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146, 102551. <https://doi.org/10.1016/j.ijhcs.2020.102551>
- Thuy, A., & Benoit, D. F. (2024). Explainability through uncertainty: Trust-worthy decision-making with neural networks. *European Journal of Operational Research*, 317(2), 330–340. <https://doi.org/10.1016/j.ejor.2023.09.009>
- Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. *International conference on machine learning*, 9690–9700.
- Van Belle, J., Guns, T., & Verbeke, W. (2021). Using shared sell-through data to forecast wholesaler demand in multi-echelon supply chains. *European Journal of Operational Research*, 288(2), 466–479. <https://doi.org/10.1016/j.ejor.2020.05.059>
- Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*, 281(3), 543–558. <https://doi.org/10.1016/j.ejor.2019.08.015>
- Verboven, S., Berrevoets, J., Wuytens, C., Baesens, B., & Verbeke, W. (2021). Autoencoders for strategic decision support [Interpretable Data Science For Decision Making]. *Decision Support Systems*, 150, 113422. <https://doi.org/10.1016/j.dss.2020.113422>
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Watson, D., O’Hara, J., Tax, N., Mudd, R., & Guy, I. (2023). Explaining predictive uncertainty with information theoretic shapley values. *Advances in Neural Information Processing Systems*, 36, 7330–7350.
- Weytjens, H., & De Weerd, J. (2022). Learning uncertainty with artificial neural networks for predictive process monitoring. *Applied Soft Computing*, 125, 109134. <https://doi.org/10.1016/j.asoc.2022.109134>
- Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., & Tingley, D. (2017). Mooc dropout prediction: How to measure accuracy? *Proceedings of the*

- fourth (2017) acm conference on learning@ scale*, 161–164. <https://doi.org/10.1145/3051457.3053974>
- Wilson, A. G. (2020). The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*. <https://doi.org/10.48550/arXiv.2001.10995>
- Yong, B. X., & Brintrup, A. (2022). Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection. *Expert Systems with Applications*, 209, 118196. <https://doi.org/10.1016/j.eswa.2022.118196>
- Yu, J., Alrajhi, L., Harit, A., Sun, Z., Cristea, A. I., & Shi, L. (2021). Exploring bayesian deep learning for urgent instructor intervention need in mooc forums. *International Conference on Intelligent Tutoring Systems*, 78–90. https://doi.org/10.1007/978-3-030-80421-3_10
- Zhang, X., & Mahadevan, S. (2020). Bayesian neural networks for flight trajectory prediction and safety assessment. *Decision Support Systems*, 131, 113246. <https://doi.org/10.1016/j.dss.2020.113246>

3

Active Learning to Guide Labeling Efforts for Question Difficulty Estimation

This chapter is based on a study published in the 2nd Workshop on Responsible Knowledge Discovery in Education (RKDE) at ECML-PKDD 2024 in Vilnius, Lithuania (Thuy et al., 2024). The versions are identical except for minor language adjustments.

Abstract

In recent years, there has been a surge in research on Question Difficulty Estimation (QDE) using natural language processing techniques. Transformer-based neural networks achieve state-of-the-art performance, primarily through supervised methods but with an isolated study in unsupervised learning. While supervised methods focus on predictive performance, they require abundant labeled data. On the other hand, unsupervised methods do not require labeled data but rely on a different evaluation metric that is also computationally expensive in practice. This work bridges the research gap by exploring active learning for QDE—a supervised human-in-the-loop approach striving to minimize the labeling efforts while matching the performance of state-of-the-art models. The active learning process iter-

actively trains on a labeled subset, acquiring labels from human experts only for the most informative unlabeled data points. Furthermore, we propose a novel acquisition function PowerVariance to add the most informative samples to the labeled set, a regression extension to the PowerBALD function popular in classification. We employ DistilBERT for QDE and identify informative samples by applying Monte Carlo dropout to capture epistemic uncertainty in unlabeled samples. The experiments demonstrate that active learning with PowerVariance acquisition achieves a performance close to fully supervised models after labeling only 10% of the training data. The proposed methodology promotes the responsible use of educational resources, makes QDE tools more accessible to course instructors, and is promising for other applications such as personalized support systems and question-answering tools.

3.1 Introduction

Question Difficulty Estimation (QDE), also known as question calibration, is a regression task that estimates a question’s difficulty directly from the question and answers’ text. It is a crucial task in personalized support tools like computerized adaptive testing (Van der Linden & Glas, 2000), which tailors questions to a student’s skill level. If the questions are too easy or too difficult, the student might lose motivation, negatively affecting their learning outcome (Wang et al., 2014).

Traditionally, QDE has been performed with manual calibration (Attali et al., 2014) and pretesting (Lane et al., 2016), which are time-consuming and expensive. Recent studies aim to address these limitations by leveraging natural language processing (NLP) techniques. The NLP approaches train machine learning models to estimate question difficulty from its text. Once trained, the models can quickly calibrate unseen questions, reducing the need for pretesting and manual calibration.

Supervised techniques dominate QDE with state-of-the-art results (Benedetto et al., 2021; Zhou & Tao, 2020) by fine-tuning the publicly available pre-trained models BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019). However, fine-tuning often requires a large labeled dataset containing tens of thousands of calibrated questions, almost impossible to collect for individual course instructors developing QDE tools on their exam data. An isolated study (Loginova et al., 2021) has delved into an unsupervised approach, relying solely on additional pre-training and evaluating pairwise difficulty. Although this approach is helpful, its performance cannot be directly compared to supervised methods and is more computationally expensive in practical implementations.

In this work, we explore *active learning* (AL) (Settles, 2009) for QDE, a data-efficient supervised approach aiming to minimize the labeling work for human annotators while matching the performance of state-of-the-art models. AL operates by iteratively training a model on an increasingly growing labeled subset by acquiring labels from an expert only for the most informative unlabeled data points. This human-in-the-loop strategy allows us to preserve the well-established supervised evaluation methods, effectively bridging the gap between the performance-driven supervised domain and the data-centric unsupervised domain. Moreover, we propose a novel acquisition function *PowerVariance* to add the most informative samples to the labeled set while limiting redundant information, a regression extension to the PowerBALD function (Kirsch et al., 2023) popular in classification. We use DistilBERT (Sanh et al., 2019) for QDE and find informative samples by applying Monte Carlo (MC) dropout (Gal & Ghahramani, 2016) to capture epistemic uncertainty over the unlabeled samples.

The proposed methodology contributes to the responsible use of educational resources by drastically reducing the labeling work, making the development of QDE tools more accessible to course instructors. The findings have positive implications for a variety of applications like personalized support tools, essay correction tools, and question-answering systems.

The remainder of the chapter is organized as follows. Section 3.2 provides an overview of related work, followed by the proposed AL methodology in Section 3.3. Experimental details are discussed in Section 3.4, with the results and discussion presented in Section 3.5. Finally, Section 3.6 concludes the chapter. The code is available in a GitHub repository.¹

3.2 Related Work

Earliest NLP research on QDE from text primarily focused on multiple-choice questions (MCQs), employing bag-of-words techniques and assessing similarities among questions, correct choices, and incorrect choices (Alsubait et al., 2013; Kurdi et al., 2017; Yaneva et al., 2018). However, these methods have been outperformed by more recent machine learning approaches.

Machine learning approaches to QDE fall into two main categories: (i) those involving distinct feature engineering and regression phases, and (ii) end-to-end neural networks (NNs). The former encompasses a wide range of features, including linguistic features, text embeddings, frequency-based features, and readability indexes. Several studies have also explored com-

¹<https://github.com/arthur-thuy/qde-active-learning>

binations of these feature techniques (Benedetto, 2023). Common machine learning regression models in this group include random forests, support vector machines, and linear regression (Benedetto et al., 2023).

End-to-end NN approaches in previous research primarily rely on Transformer models (Vaswani et al., 2017), which can be either supervised or unsupervised. Transformers are attention-based NNs pre-trained on a large corpus of text. This generally yields superior performance with shorter training times compared to training NNs from scratch, leveraging the pre-existing knowledge of the pre-trained model.

Supervised estimation to QDE is most prevalent in the literature (Cheng et al., 2019; Qiu et al., 2019; Tong et al., 2020). Fine-tuning the publicly available pre-trained models BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) on the task of QDE gives state-of-the-art results (Benedetto et al., 2021; Zhou & Tao, 2020) and has been shown to outperform other approaches using traditional NLP-derived features (Benedetto, 2023).

Unsupervised estimation, aiming to avoid relying on labeled data entirely, has received comparatively less attention. One study (Loginova et al., 2021) estimates question difficulty by leveraging the epistemic uncertainty in question answering models as an indicator of human-perceived difficulty. This approach involves additional pre-training without fine-tuning, making it independent of labeled data. While helpful in estimating difficulty, its performance cannot be directly compared to supervised estimation as it evaluates pairwise difficulty. Moreover, it poses computational challenges in practice because numerous pairwise evaluations are required to determine an overall difficulty ranking of unseen questions.

3.3 Methodology

3.3.1 Active Learning

AL (Settles, 2009) is a human-in-the-loop technique for achieving data efficiency. Instead of collecting and labeling a large dataset before training, which is time-consuming and expensive, AL iteratively acquires labels from an expert annotator only for the most informative data points from a pool of unlabeled data. After each acquisition step, the newly labeled points are added to the training set, and the model is retrained. This process is repeated until reaching a desired level of accuracy or until the labeling budget is exhausted, aiming to minimize the labeling work of human annotators. Figure 3.1 provides a visual overview of the AL workflow, employing pool-based sampling as described.

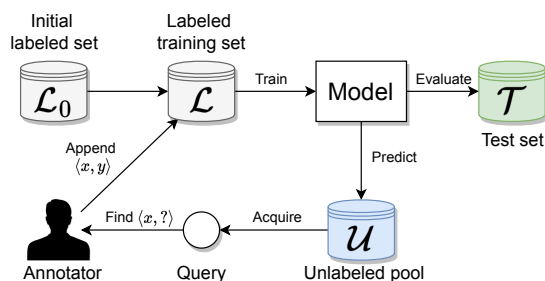


Figure 3.1: Active learning workflow with pool-based sampling. Active learning iteratively trains on a subset of labeled data and acquires labels from an expert annotator for samples in the unlabeled pool. Adapted from Settles (2009).

In AL, the informativeness of new points is assessed by an acquisition function. The acquisition function typically relies on epistemic uncertainty over unlabeled data, which can be obtained with approximate Bayesian inference techniques like MC dropout (Gal & Ghahramani, 2016) or with ensembling techniques (Lakshminarayanan et al., 2017; Thuy & Benoit, 2024a, 2024b). Epistemic uncertainty represents uncertainty in the model parameters and is naturally high in regions of the input space with few training observations (Der Kiureghian & Ditlevsen, 2009), precisely the observations we want to add to the labeled set. For classification tasks, a commonly used acquisition scoring function is Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011), which estimates the epistemic uncertainty by measuring the variability among samples of the predictive distribution. Data points maximize this acquisition function when the model assigns the highest predicted probability to a different class in each sample. For regression tasks, the epistemic uncertainty is estimated by the Variance among predictive samples (Settles, 2009). Similarly, data points score high on this acquisition function when the model’s output varies strongly across the samples.

3.3.2 Monte Carlo Dropout Uncertainty

Uncertainty in predictions can arise from two different sources: aleatoric and epistemic uncertainty (Der Kiureghian & Ditlevsen, 2009). Aleatoric uncertainty refers to the notion of randomness and is related to the data-measurement process. This uncertainty is irreducible even if more data is collected. Epistemic uncertainty accounts for uncertainty in the model parameters. In contrast to aleatoric uncertainty, collecting more data can reduce epistemic uncertainty. As such, it is interesting for acquisitions func-

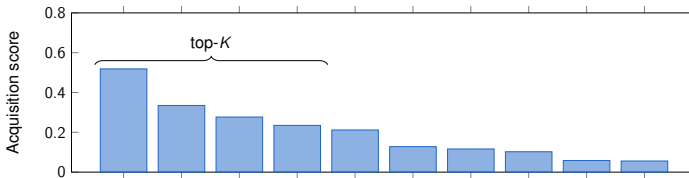


Figure 3.2: **Top- K acquisition toy example.** Acquisition scores for each unlabeled pool point are ordered and the top- K points are selected.

tions to select the unlabeled samples with the largest epistemic uncertainty.

We assume a regression task with inputs \mathbf{X} , labels \mathbf{Y} , and a discriminative regressor $p(\mathbf{y} \mid \mathbf{x})$. For the Bayesian MC dropout models, we further assume a probability distribution over the parameters, $p(\boldsymbol{\theta})$, and we have $p(\mathbf{y} \mid \mathbf{x}) = \mathbb{E}_{p(\boldsymbol{\theta})}[p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})]$. In a NN regressor, the output \mathbf{y} represents the mean $\mu_{\mathbf{x}}$ of the conditional probability distribution $\mathcal{N}(\mu_{\mathbf{x}}, \sigma = 1)$, for some input point \mathbf{x} . The standard NN regressor only outputs a single $\mu_{\mathbf{x}}$, hence does not capture any uncertainty. With MC dropout, multiple estimates for $\mu_{\mathbf{x}}$ are obtained and the variance over these estimates is an approximation for the epistemic uncertainty in data point \mathbf{x} .

In MC dropout (Gal & Ghahramani, 2016), dropout is not only applied at training time but also at test time. Multiple forward passes are performed, each time randomly dropping units and getting another thinned dropout variant of the NN. As such, it can be seen as an implicit ensemble method where each sample corresponds to an ensemble member. The various samples approximate the true posterior predictive distribution, enabling it to estimate the epistemic uncertainty in a data point.

3.3.3 PowerVariance Acquisition

In practical AL applications, instead of single data points, batches of data points are selected in each acquisition step to minimize the frequency of model retraining and expert involvement. A common heuristic involves selecting the top- K highest-scoring points from an acquisition scheme designed for single-point selection, i.e., top- K acquisition (Kirsch et al., 2019) (Figure 3.2). However, this method overlooks interactions between points within an acquisition batch because individual points are scored independently. For example, if the most informative point is duplicated in the pool set, all instances will be acquired, which is wasteful.

To address this issue, acquisition functions designed explicitly for batch acquisition with NN classifiers have been developed, such as BatchBALD (Kirsch et al., 2019). These methods improve over top- K acquisition by

accounting for the interaction between points but are computationally expensive. To limit the computational burden, Kirsch et al. (2023) propose to stochastically acquire points following a power distribution determined by the single-acquisition scores. Intuitively, points with high acquisition scores are more likely to be sampled. For example, for BALD, the method is referred to as PowerBALD, demonstrating equal performance to state-of-the-art batch acquisition functions like BatchBALD while requiring significantly less computational resources.

The stochastic acquisition strategy (Kirsch et al., 2023) assumes that as new samples are selected in a batch, future acquisition scores differ from the current scores by a perturbation. This perturbation is modeled as Gumbel-distributed noise for two reasons.

First, to select the k -th point in the acquisition batch of size K , it is important to consider how much additional information (i.e., increase in acquisition scores) the still-to-be-selected $K - k$ points will provide. As such, the stochastic strategy models the maximum future increase in acquisition scores over all possible candidate points to complete the batch. Empirically, acquisition scores are similar to a truncated exponential distribution, with different rate parameters at each AL step. The maximum over sums of such random variables is empirically shown to follow a Gumbel distribution (Kirsch et al., 2023).

Second, the Gumbel distribution is also mathematically convenient. The Gumbel-Top- K trick (Kool et al., 2019) shows that taking the highest-scoring points from a distribution perturbed with Gumbel noise is equivalent to sampling from a softmax distribution without replacement. Building on this, perturbing the log-scores with Gumbel noise results in sampling from a power distribution. Power acquisition assumes that scores are non-negative and uninformative points should be avoided, a sensible approach for AL.

We propose to extend this approach to regression settings, which is currently underinvestigated, resulting in a PowerVariance acquisition function. Similar to BALD, the Variance scoring function is non-negative, with zero variance indicating an uninformative sample due to no expected information gain. Consequently, the Variance function should also couple well with power acquisition, mirroring the success seen with BALD and PowerBALD.

More formally, for each candidate pool index i , the Variance score is

$$s_{Var}(i) = \text{Var}[p(\mathbf{y} \mid \mathbf{x}_i, \boldsymbol{\theta})]. \quad (3.1)$$

The PowerVariance score is the perturbation of the log Variance score with

Gumbel-distributed noise $\epsilon_i \sim \text{Gumbel}(0; \beta^{-1})$

$$s_{PowerVar}(i) = \log s_{Var}(i) + \epsilon_i. \quad (3.2)$$

Following the Gumbel-Top- K trick (Kool et al., 2019), taking the top- K points from $s_{PowerVar}$ is equivalent to sampling without replacement from the distribution $p_{PowerVar}$

$$p_{PowerVar}(i) \propto s_{Var}(i)^\beta \quad (3.3)$$

where $\beta \geq 0$ is a *coldness* parameter. Note that the coldness parameter β is different but similar to the *temperature* parameter $T = 1/\beta$ often used in text-generation with language models. For $\beta \rightarrow \infty$, this strategy converges towards top- K acquisition as it is more likely to only sample points with a high score. For $\beta \rightarrow 0$, it converges towards uniform acquisition because it is more likely to also sample points with a low score.

3.4 Experiments

3.4.1 Data

RACE++ (Liang et al., 2019) is a dataset of reading comprehension MCQs, built by merging the original RACE dataset (Lai et al., 2017) with the RACE-C dataset (Liang et al., 2019). Each question comprises a reading passage, a stem, and four possible answer options, one of them being correct. Each question has one out of three difficulty levels (0, 1, 2), which we consider as the gold standard for QDE. The difficulty levels correspond to middle school, high school, and university-level questions; the dataset is imbalanced, with a distribution of 25%, 62%, and 13% respectively. Note that the dataset labels are all available; the labels are hidden and revealed once requested by the acquisition function. The training split contains 100,568 questions, while the validation and test splits contain 1000 and 5642 questions, respectively. There are no reading passages shared across the splits.

3.4.2 Model Architecture

We fine-tune the publicly available pre-trained model DistilBERT on the task of QDE. DistilBERT is a language model obtained by distilling BERT, i.e., compressing BERT by training a small model to reproduce its full output distribution (Hinton et al., 2015). The authors of Benedetto et al. (2023) find that DistilBERT achieves comparable performance to BERT on QDE

while using approximately half the parameter count. Limiting the computational expense is important in AL as the model needs to be fine-tuned over multiple iterations.

To adapt DistilBERT for QDE, we stack a fully connected hidden layer on top of the pre-trained language model, followed by the output layer. The regression output layer has one unit, with its weights initialized randomly. During fine-tuning, both the weights of the output head and the pre-trained model are updated. We follow the input encoding of Benedetto (2023) and concatenate the question and the text of all the possible answer choices in a single sentence, divided by separator tokens. This configuration has demonstrated slight improvements over using no answer choice at all or only the correct answer.

Following previous research (Benedetto, 2023), we handle QDE on the RACE++ dataset as a discrete regression problem. The QDE model is trained as a regression model and outputs a continuous difficulty, which is then converted to the closest discrete level with simple thresholds. As evaluation metric, we compute the root mean squared error (RMSE) between the discrete predictions and discrete difficulty levels because of its consideration for the order of difficulty levels. We refer to this metric as “discrete RMSE”.

3.4.3 Active Learning Setup

The AL process starts with an initial labeled dataset of 500 observations, randomly selected from the training set and following the training set distribution (i.e., 25%/62%/13%). In each iteration, the model is fine-tuned for 10 training epochs and the parameters giving the best validation performance are saved. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with learning rate $2e-5$ and batch size 64.

Subsequently, the model is evaluated on a random subset of the unlabeled pool containing 5000 observations, from which 100 observations are selected for labeling and added to the training set. As demonstrated by Atighehchian et al. (2022), using a random subset instead of the entire pool minimally impacts predictive performance while being more computationally efficient. The configurations with MC dropout use 10 MC samples to calculate epistemic uncertainty. Before training on the new training set, the model weights are re-initialized to their original state. This involves using pre-trained weights for the base model and randomly initialized weights for the output layer with He initialization (He et al., 2015), effectively mitigating the risk of getting stuck in a poor local minimum. Table 3.1 displays all hyperparameter settings for the model and AL setup.

Table 3.1: *Hyperparameter settings.*

<i>(a) Model</i>		<i>(b) Active learning</i>	
Hyperparameter	Value	Hyperparameter	Value
Model name	DistilBERT	Dataset name	RACE++
Learning rate	2e-5	Data size	100,568/ train/val/test
Optimizer	AdamW	Initial labeled set size	1000/5642 500
Weight decay	0.05	Acquisition size	100
Loss function	MSE	Pool subset size	5000
Training epochs	10	Final labeled set size	10,000
Train batch size	64	MC samples	10
Eval batch size	256		
Dropout rate	0.1		
Warmup ratio	0.1		
Sequence length	256		

In the experiments, we compare three AL configurations: (i) Uniform acquisition with a standard NN, (ii) top- K Variance acquisition with an MC dropout NN, and (iii) PowerVariance acquisition with an MC dropout NN. For PowerVariance acquisition, we follow Kirsch et al. (2023) and set $\beta = 1$ to limit the number of hyperparameters. Note that Uniform acquisition is computationally cheaper because it does not predict on the unlabeled pool, instead it randomly selects observations for labeling.

Additionally, we investigate the performance of three baselines: (i) Random, (ii) Majority, and (iii) Supervised. The Random baseline randomly predicts a difficulty level, the Majority baseline consistently predicts level 1 (the most prevalent level in the training set), and the Supervised baseline fine-tunes a model on the fully labeled training set. Consequently, the Random and Majority baselines serve as performance lower bounds, while the Supervised baseline sets an upper bound.

The experiments are implemented in PyTorch (Ansel et al., 2024) using the BaaL (Atighehchian et al., 2022) and HuggingFace (Wolf et al., 2020) packages, executed on an NVIDIA RTX A5000 GPU. The results are averaged over five independent runs with random seeds, with a total runtime of 90 hours.

3.5 Results and Discussion

Section 3.5.1 explores the AL results, while Section 3.5.2 provides a more detailed analysis of the behavior of the acquisition functions.

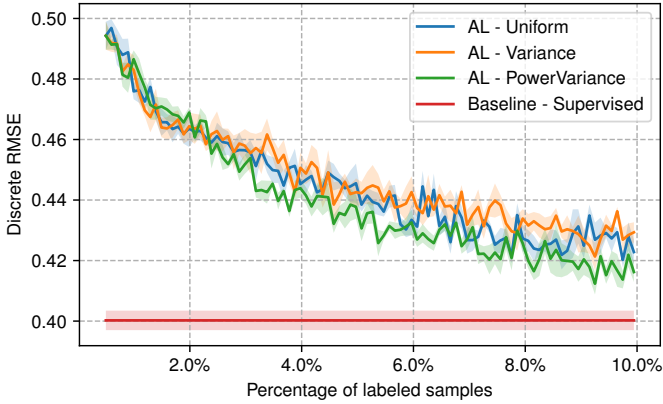


Figure 3.3: *Discrete RMSE as a function of the labeled dataset size. PowerVariance acquisition outperforms Uniform and Variance acquisition by achieving the lowest discrete RMSE scores as AL progresses. After labeling 10% of the data, its performance is close to the fully supervised model.*

3.5.1 Predictive Performance

Figure 3.3 shows the discrete RMSE in relation to the training dataset size. For AL configurations, values to the lower left indicate better performance. The Supervised baseline’s performance is represented by a horizontal line, where lower is better. The Random baseline achieves a discrete RMSE of 1.026 and the Majority baseline achieves 0.616. Note that these baselines are not shown in Figure 3.3 to avoid an excessively large vertical axis, which complicates interpretation.

The findings reveal that fine-tuning on the initial labeled set (500 observations; 0.5%) performs exactly in between the Majority and Supervised baselines. As the labeled training set expands, we observe a decrease in discrete RMSE across all AL configurations. This finding demonstrates that DistilBERT performs well with limited labeled data, consistent with previous research (Sun et al., 2019) using the non-distilled BERT model.

Variance acquisition disappoints and performs on par with Uniform acquisition. This result is surprising given that Variance acquisition uses MC dropout to quantify epistemic uncertainty over the unlabeled pool points. As such, naively selecting the top- K highest scoring-points does not yield improved results.

In contrast, PowerVariance acquisition outperforms both Uniform and Variance acquisition, achieving the lowest discrete RMSE score from 2% of labeled samples onwards. Although PowerVariance’s RMSE advantage

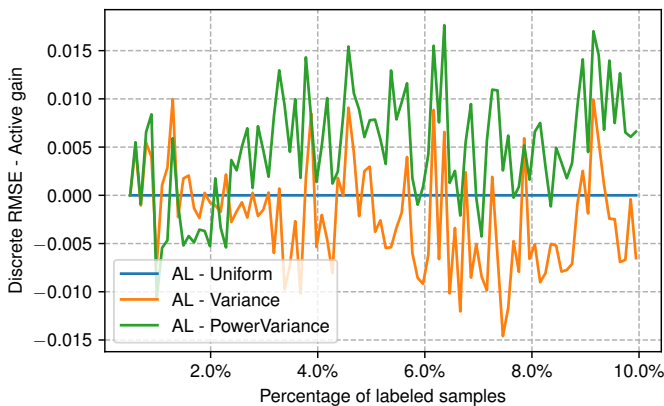


Figure 3.4: *Active gain over Uniform acquisition as a function of the labeled dataset size. Variance acquisition performs on par with passive learning, while PowerVariance offers an active gain of 0.01 discrete RMSE.*

over Random acquisition appears minimal due to the line curves being close, a substantial number of labeled questions is needed to overcome this advantage. After collecting a labeled set containing 10% of the available samples, PowerVariance reaches a discrete RMSE score only 5% higher than training on 100% of the training data.

AL with Uniform acquisition is often referred to as *passive learning*, as samples are randomly selected from the unlabeled pool. Figure 3.4 illustrates the active gain in discrete RMSE, highlighting the performance differences of Variance and PowerVariance over Uniform acquisition. Positive values denote an advantage, while negative values indicate a disadvantage, enabling relative comparisons among acquisition functions.

From 2% of labeled data onwards, PowerVariance exhibits a positive active gain, averaging around 0.01 discrete RMSE. In contrast, Variance acquisition does not offer advantages over passive learning. The next subsection delves deeper into the acquisition functions, examining the reasons behind the performance differences.

3.5.2 Acquisition Behavior

To better understand how the acquisition functions behave, we visualize the distribution of difficulty levels in the labeled set as training progresses (see Figure 3.5). Initially, the labeled set is randomly sampled from the pool following a 25%/62%/13% distribution for levels 0, 1, and 2, respectively. Due to the small sample size (500 samples), slight deviations from this

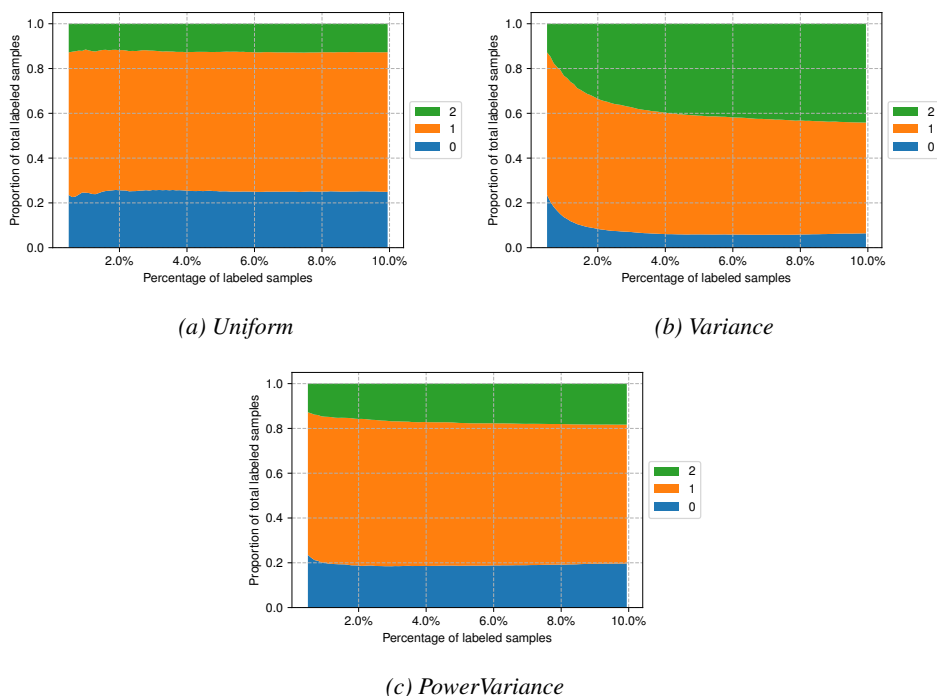


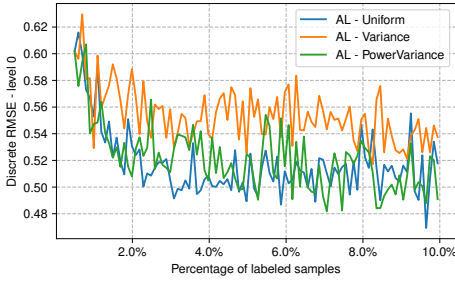
Figure 3.5: **Distribution of difficulty levels in the labeled set as a function of the labeled dataset size, per acquisition function.** Similar to Variance, PowerVariance selects more level 2 observations but does not neglect level 0 samples.

distribution are possible.

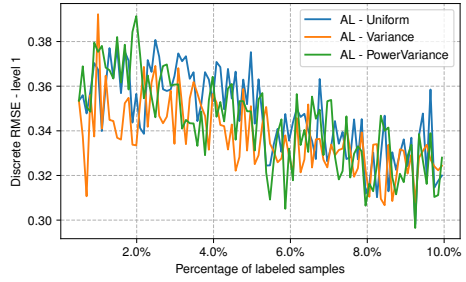
As expected, Uniform acquisition causes minimal changes in the level distribution because samples are randomly selected from the unlabeled pool. In contrast, Variance acquisition exhibits a distinctive pattern, selecting many level 2 observations and few level 0 instances. The proportion of level 2 samples increases from 13% to 45%, while level 0 samples decrease from 25% to merely 6%. These findings partially align with previous studies (Atighehchian et al., 2020) suggesting that top- K strategies using epistemic uncertainty target underrepresented classes. Variance acquisition indeed prioritizes sampling from the most underrepresented class (level 2) but does this primarily at the expense of level 0 observations, rather than the majority class (level 1).

PowerVariance exhibits behavior that falls between Uniform and Variance acquisition. Like Variance acquisition, it selects more level 2 observations, increasing their proportion from 13% to 19%, while only slightly reducing the level 0 proportion from 25% to 20%. As such, it is a less

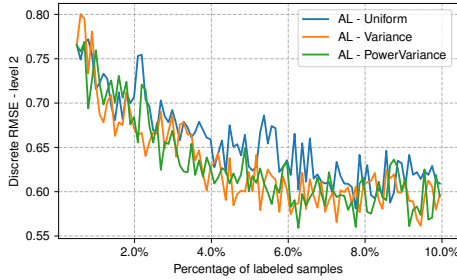
ACTIVE LEARNING TO GUIDE LABELING EFFORTS FOR QUESTION DIFFICULTY ESTIMATION



(a) Level 0 (middle school)



(b) Level 1 (high school)



(c) Level 2 (university)

Figure 3.6: Discrete RMSE as a function of the dataset size, per individual difficulty level. Variance and PowerVariance surpass Uniform acquisition on level 2, but Variance underperforms on level 0.

aggressive approach compared to Variance acquisition.

Moreover, we analyze the impact of the acquisition strategies on the predictive performance for each difficulty level individually. Figure 3.6 displays the discrete RMSE performance per difficulty level.

For level 1 questions (Figure 3.6b), all acquisition functions have comparable performance as the line graphs overlap. However, we observe performance differences on level 0 and level 2 questions.

For level 0 questions (Figure 3.6a), Variance acquisition performs poorly as the orange curve is notably higher than the other curves. This poor performance is a direct consequence of neglecting level 0 observations during acquisition.

For level 2 questions (Figure 3.6c), Uniform acquisition performs worst. Level 2 observations are the most difficult to estimate and the most underrepresented in the initial labeled set. Variance and PowerVariance sample a large number of these questions and therefore achieve good RMSE scores, lower than Uniform acquisition. It is also worth noting that the per-

formance of Variance and PowerVariance is very similar, although Variance samples a much higher proportion of level 2 questions (45%) than PowerVariance (19%).

For each difficulty level, PowerVariance acquisition performs on par or better than Uniform and Variance acquisition. It leverages epistemic uncertainty to sample more from underrepresented level 2 questions which are most challenging to estimate, whereas Uniform acquisition naively samples at random. Furthermore, PowerVariance recognizes redundant uncertainty information in level 2 questions and instead samples from level 0 questions, whereas Variance neglects level 0 questions, significantly hampering its performance.

3.6 Conclusion

This work explores AL for QDE, a supervised approach aiming to minimize the labeling effort for human annotators while matching the performance of state-of-the-art models. By using a human-in-the-loop method, it bridges the gap between the performance-driven supervised domain and the data-centric unsupervised domain. Additionally, we introduce a novel acquisition function PowerVariance, which leverages epistemic uncertainty from unlabeled samples obtained through MC dropout to identify the most informative data points. Unlike conventional Variance acquisition, PowerVariance is designed to limit redundant information in a batch of samples.

Experimental results indicate that the proposed PowerVariance acquisition outperforms both Uniform and Variance acquisition. It effectively selects observations from the minority difficulty level 2 for labeling and does not neglect level 0 questions, an issue observed with Variance acquisition. We see no reason for practitioners to consider the flawed top- K Variance acquisition. Even with only 10% of the training data labeled, AL with PowerVariance achieves good performance, only 5% higher discrete RMSE than the model trained on 100% of the training data.

This methodology promotes the responsible use of educational resources by significantly reducing the labeling work for educational professionals while maintaining predictive performance. Consequently, it makes QDE tools more accessible to course instructors who might otherwise be demotivated by the large number of calibrated questions required.

The study is potentially limited by the small number of coarse difficulty levels. Course instructors are often reluctant to share exam questions, making it challenging to find datasets with more realistic difficulty levels. Future research can explore more fine-grained settings with more closely spaced difficulty levels. The inability to use public datasets highlights the

relevance of active learning strategies for course instructors when labeling exam questions. Furthermore, adding difficulty levels may introduce class imbalance, a scenario where PowerVariance performs strongly.

The proposed AL approach holds promise for diverse applications such as personalized support tools, essay correction tools, and question-answering systems. It can easily be adapted to alternative pre-trained language models and datasets, as MC dropout works on any architecture that uses dropout. For models not employing dropout, ensembles of NNs can provide epistemic uncertainty, enabling similar AL strategies.

Reflections and Clarifications

Computational Cost

Reflecting on the experimental results, it becomes evident that the computational cost of active learning is substantially higher than that of conventional supervised learning. By iteratively selecting samples for human annotation, active learning effectively trades human effort for computational expense. This trade-off highlights a broader infrastructural challenge discussed in Chapter 1: institutions must decide between investing in on-premise computing resources, requiring significant upfront costs, or leveraging pay-per-use cloud services, which are often more accessible to smaller schools. Looking ahead, we anticipate that the relevance of active learning will continue to grow, as the cost of human labor is expected to rise, while the cost of computation will likely decline due to ongoing large-scale investments in data center infrastructure.

Bias in Active Learning

A key consideration in active learning is sampling bias, i.e., the bias that arises when the labeled subset is distributed differently from the overall dataset. As noted by Settles (2009), sampling bias is an inherent characteristic of nearly all active learning strategies. To some extent, such bias is even desirable, as selectively oversampling minority classes can improve predictive performance. However, the experimental results in Chapter 3 show that Top- K acquisition may excessively favor a minority class (level 2), to the point where the central class (level 0) is underrepresented, ultimately reducing overall performance. In contrast, the proposed PowerVariance approach avoids this issue by striking a balance between Top- K and Uniform acquisition, achieving superior performance while mitigating excessive sampling bias.

Data Annotation with LLMs

Given the substantial cost associated with human annotation, an alternative approach to reduce the amount of human labor is to employ an LLM as an annotator instead of a human expert. This strategy effectively transforms the paradigm from *human-in-the-loop* to *LLM-in-the-loop*.

A related concept is self-training (Amini et al., 2025), a variant of active learning in which the same model used for training also generates labels, commonly referred to as pseudo-labeling. However, relying on the same model for both tasks can amplify bias in the labeled data, compounding the inherent sampling bias already present in standard active learning. By contrast, the LLM-in-the-loop approach offers a more robust alternative: because the generative LLM is pre-trained on diverse data sources and exhibits different error patterns than the task-specific machine learning model, it can provide complementary annotations. Moreover, interacting with a generative LLM closely mirrors the process of engaging with a human annotator, making this approach both intuitive and practical in educational settings.

PowerVariance Acquisition Function

The PowerVariance acquisition function is not limited to NNs. It is a general “query by committee” approach (Settles, 2009) that operates on a set of model predictions to quantify the degree of disagreement, i.e., an indicator of epistemic uncertainty. Consequently, PowerVariance can also be applied to other model families, including Bayesian methods such as Gaussian Processes and ensembling approaches like Random Forests and Gradient Boosting (e.g., CatBoost, as discussed in the reflections of Chapter 2).

Neural Network Settings

In designing the experiments and tuning hyperparameters, several decisions have been made to reduce computational cost.

Hyperparameters for the NN have been obtained with manual hyperparameter tuning on Uniform acquisition, minimizing the validation RMSE after labeling 10% of the training questions. Hyperparameters include the batch size, learning rate, weight decay, and dropout rate. To further limit computational overhead, the same hyperparameter configuration was applied to both active learning setups using MC Dropout (i.e., Top- K Variance and PowerVariance), as the underlying NN architecture remains identical, except that dropout is also enabled at test time. Hyperparameter tuning in active learning is inherently more complex than in standard super-

vised learning, since it requires evaluating a sequence of validation curves rather than a single one, making the process less straightforward for practitioners.

In the experiments, MC Dropout used a sample size of 10, which is considerably smaller than the configuration used in Chapter 2, due to the increased computational demands of the iterative active learning process. Nonetheless, the benchmarking study by Ovadia et al. (2019) reported diminishing returns beyond a sample size of 5, suggesting that using 10 samples should not adversely affect the quality of epistemic uncertainty estimates.

Following Kirsch et al. (2023), the β parameter in the Power acquisition function was fixed to 1.0 to further control computational expense. However, since varying β may substantially influence the PowerVariance scores, tuning β presents a promising direction for future performance improvements.

Finally, the AL experiments were run until 10% of the data had been labeled, as performance curves typically plateaued around this point. Ideally, the process would continue until full labeling (100%), but this was computationally infeasible for a large dataset such as RACE++ (approximately 100,000 observations).

Acknowledgments

This study was supported by the Research Foundation Flanders (FWO) (grant number 1S97022N).

References

- Alsubait, T., Parsia, B., & Sattler, U. (2013). A similarity-based theory of controlling mcq difficulty. *2013 second international conference on e-learning and e-technologies in education (ICEEE)*, 283–288. <https://doi.org/10.1109/ICeLeTE.2013.6644389>
- Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., & Maximov, Y. (2025). Self-training: A survey. *Neurocomputing*, *616*, 128904.
- Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., . . . Chintala, S. (2024). PyTorch 2: Faster

- Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. <https://doi.org/10.1145/3620665.3640366>
- Atighehchian, P., Branchaud-Charron, F., Freyberg, J., Pardinias, R., Schell, L., & Pearse, G. (2022). Baal, a bayesian active learning library.
- Atighehchian, P., Branchaud-Charron, F., & Lacoste, A. (2020). Bayesian active learning for production, a systematic study and a reusable library. *arXiv preprint arXiv:2006.09916*. <https://doi.org/10.48550/arXiv.2006.09916>
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series, 2014(2)*, 1–8. <https://doi.org/10.1002/ets2.12042>
- Benedetto, L. (2023). A quantitative study of nlp approaches to question difficulty estimation. *International Conference on Artificial Intelligence in Education*, 428–434. https://doi.org/10.1007/978-3-031-36336-8_67
- Benedetto, L., Aradelli, G., Cremonesi, P., Cappelli, A., Giussani, A., & Turrin, R. (2021). On the application of transformers for estimating the difficulty of multiple-choice questions from text. *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*, 147–157.
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1–37. <https://doi.org/10.1145/3556538>
- Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., Chen, Y., Ma, H., & Hu, G. (2019). Dirt: Deep learning enhanced item response theory for cognitive diagnosis. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2397–2400. <https://doi.org/10.1145/3357384.3358070>
- Der Kiureghian, A., & Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2), 105–112. <https://doi.org/10.1016/j.strusafe.2008.06.020>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings*

- of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *international conference on machine learning*, 1050–1059.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*, 1026–1034. <https://doi.org/10.1109/ICCV.2015.123>
- Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. <https://doi.org/10.48550/arXiv.1503.02531>
- Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*. <https://doi.org/10.48550/arXiv.1112.5745>
- Kirsch, A., Farquhar, S., Atighehchian, P., Jesson, A., Branchaud-Charron, F., & Gal, Y. (2023). Stochastic batch acquisition: A simple baseline for deep active learning. *Trans. Mach. Learn. Res.*, 2023. <https://openreview.net/forum?id=vcHwQyNBjW>
- Kirsch, A., Van Amersfoort, J., & Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32.
- Kool, W., Van Hoof, H., & Welling, M. (2019). Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. *International Conference on Machine Learning*, 3499–3508.
- Kurdi, G., Parsia, B., & Sattler, U. (2017). An experimental evaluation of automatically generated multiple choice questions from ontologies. *OWL: Experiences and Directions–Reasoner Evaluation: 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, Italy, November 20, 2016, Revised Selected Papers 13*, 24–39. https://doi.org/10.1007/978-3-319-54627-8_3
- Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017, September). RACE: Large-scale ReAding comprehension dataset from examinations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017*

- conference on empirical methods in natural language processing* (pp. 785–794). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1082>
- Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- Lane, S., Raymond, M. R., Haladyna, T. M., et al. (2016). *Handbook of test development* (Vol. 2). Routledge New York, NY.
- Liang, Y., Li, J., & Yin, J. (2019). A new multi-choice reading comprehension dataset for curriculum learning. *Asian Conference on Machine Learning*, 742–757.
- Loginova, E., Benedetto, L., Benoit, D., & Cremonesi, P. (2021). Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. *RANLP 2021*, 846–855. https://doi.org/https://doi.org/10.26615/978-954-452-072-4_097
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- Qiu, Z., Wu, X., & Fan, W. (2019). Question difficulty prediction for multiple choice problems in medical exams. *Proceedings of the 28th acm international conference on information and knowledge management*, 139–148. <https://doi.org/10.1145/3357384.3358013>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. <https://doi.org/10.48550/arXiv.1910.01108>
- Settles, B. (2009). Active learning literature survey.
- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, 194–206.
- Thuy, A., & Benoit, D. F. (2024a). Explainability through uncertainty: Trustworthy decision-making with neural networks. *European*

- Journal of Operational Research*, 317(2), 330–340. <https://doi.org/10.1016/j.ejor.2023.09.009>
- Thuy, A., & Benoit, D. F. (2024b). Fast and reliable uncertainty quantification with neural network ensembles for industrial image classification. *Annals of Operations Research*, 1–27. <https://doi.org/https://doi.org/10.1007/s10479-024-06440-4>
- Thuy, A., Loginova, E., & Benoit, D. F. (2024). Active learning to guide labeling efforts for question difficulty estimation. *arXiv preprint arXiv:2409.09258*. <https://doi.org/10.48550/arXiv.2409.09258>
- Tong, H., Zhou, Y., & Wang, Z. (2020). Exercise hierarchical feature enhanced knowledge tracing. *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21*, 324–328. https://doi.org/10.1007/978-3-030-52240-7_59
- Van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Springer. <https://doi.org/10.1007/0-306-47531-6>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, Q., Liu, J., Wang, B., & Guo, L. (2014). A regularized competition model for question difficulty estimation in community question answering services. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1115–1126. <https://doi.org/10.3115/v1/D14-1118>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- Yaneva, V., et al. (2018). Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 389–398. <https://doi.org/10.18653/v1/W18-0548>
- Zhou, Y., & Tao, C. (2020). Multi-task bert for problem difficulty prediction. *2020 international conference on communications, informa-*

tion system and computer engineering (cisce), 213–216. <https://doi.org/10.1109/CISCE50729.2020.00048>

4

Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and OrderedLogitNN

This chapter is based on a study published in the 2nd Workshop on Automated Evaluation of Learning and Assessment Content (EvalLAC) at AIED 2025 in Palermo, Italy (Thuy et al., 2025). The versions are identical except for minor language adjustments.

Abstract

Recent years have seen growing interest in Question Difficulty Estimation (QDE) using natural language processing techniques. Question difficulty is often represented using discrete levels, framing the task as ordinal regression due to the inherent ordering from easiest to hardest. However, the literature has neglected the ordinal nature of the task, relying on classification or discretized regression models, with specialized ordinal regression methods remaining unexplored. Furthermore, evaluation metrics are tightly coupled to the modeling paradigm, hindering cross-study comparability. While some metrics fail to account for the ordinal structure of difficulty levels, none adequately address class imbalance, resulting in biased performance assessments. This study addresses these limitations by benchmark-

ing three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced Discrete Ranked Probability Score (DRPS), a novel metric that jointly captures ordinality and class imbalance. In addition to using popular ordinal regression methods, we propose OrderedLogitNN, extending the ordered logit model from econometrics to neural networks. We fine-tune BERT on the RACE++ and ARC datasets and find that OrderedLogitNN performs considerably better on complex tasks. The balanced DRPS offers a robust and fair evaluation metric for discrete-level QDE, providing a principled foundation for future research.

4.1 Introduction

Question Difficulty Estimation (QDE), also known as question calibration, aims to predict a question’s difficulty directly from its textual content and its answer options. This task plays a central role in personalized learning tools such as computerized adaptive testing (Van der Linden & Glas, 2000) and dynamic online learning platforms, which aim to present questions aligned with a learner’s proficiency level. Selecting questions that are either too easy or excessively difficult can reduce student motivation and hinder learning outcomes (Wang et al., 2014). Reliable estimation of question difficulty is therefore essential.

Traditionally, QDE has relied on manual calibration (Attali et al., 2014) or pretesting (Lane et al., 2016), both of which are time-consuming and costly. To address these limitations, research has explored the use of natural language processing (NLP) techniques. These approaches train machine learning models to infer difficulty from the question text, allowing for rapid and scalable calibration of new questions without the need for manual intervention.

Difficulty levels in QDE are represented either as continuous scores or as discrete categories, with discrete levels being attractive for their ease of use. While continuous difficulty estimation is generally framed as a regression task, discrete-level QDE is essentially an ordinal regression problem, given the inherent ordering of difficulty levels from easiest to hardest.

However, the discrete-level QDE literature has neglected the ordinal nature of the task. Instead, existing work relies exclusively on classification and discretized regression methods, both of which are oversimplifications of the problem structure. Classification models disregard ordinal relationships altogether, and while discretized regression methods preserve ordinality, they implicitly assume equal spacing between levels—an assumption often violated in real-world data. As such, specialized ordinal regression techniques remain unexplored in this context. Moreover, no prior studies

have systematically compared these competing approaches. Compounding the issue, studies typically only report performance metrics aligning with their chosen modeling paradigm, making cross-study comparisons difficult. The metrics also fail to account for class imbalances, a prevalent issue in this setting. As a result, there is no consensus on the most effective evaluation metric or modeling approach for discrete-level QDE.

This study addresses the literature gaps in discrete-level QDE by proposing the balanced Discrete Ranked Probability Score (DRPS), a novel evaluation metric that jointly captures ordinality and class imbalance. It also provides a direct way to compare deterministic predictions to probabilistic ones, which are especially valuable for downstream decision-making. We benchmark three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced DRPS. Moreover, we propose a novel ordinal regression model OrderedLogitNN, extending the ordered logit model from econometrics to neural networks (NNs). Our work is the first to (i) introduce the balanced DRPS metric, (ii) compare classification and discretized regression models for QDE, and (iii) investigate specialized ordinal regression techniques, including the novel OrderedLogitNN. We conduct experiments by fine-tuning the Transformer model BERT on the RACE++ and ARC datasets.

The remainder of the chapter is structured as follows. Section 4.2 reviews related work. Section 4.3 introduces the balanced DRPS metric. Section 4.4 discusses the novel OrderedLogitNN model and Section 4.5 outlines existing methods for ordinal tasks. Section 4.6 describes our experimental setup, and Section 4.7 presents the results and discussion. We conclude in Section 4.8. The source code is available on GitHub.¹

4.2 Related Work

Building on the survey by Benedetto et al. (2023), we further investigate studies in QDE that utilize datasets with discrete difficulty levels. Question difficulty is defined using one of three main approaches: (i) Classical Test Theory (CTT) (Hambleton & Jones, 1993), (ii) Item Response Theory (IRT) (Hambleton, 1991), and (iii) manual calibration. In the case of manual calibration with expert annotators, difficulty is almost exclusively assigned in discrete levels due to its ease of use. While difficulty scores derived from CTT and IRT are continuous by nature, they are often discretized in practical applications to facilitate interpretation. Table 4.1 provides an overview of related work in discrete-level QDE.

¹<https://github.com/arthur-thuy/qde-ordinality>

Table 4.1: Related work on discrete-leveled QDE.

Paper	Year	Difficulty	Model output format			Metric
			Regression	Classification	Ordinal	
Hsu et al. (2018)	2018	IRT	—	✓	—	(Adjacent) accuracy
Yang and Suyong (2018)	2018	Manual	—	✓	—	Accuracy
Fang et al. (2019)	2019	Manual	—	✓	—	Accuracy
Lin et al. (2019)	2019	Manual, IRT	✓	—	—	(Adjacent) accuracy
Zhou and Tao (2020)	2020	CTT	—	✓	—	Accuracy, F_1 -score
Loginova et al. (2021)	2021	Manual	—	✓	—	Accuracy
Benedetto (2023)	2023	Manual	✓	—	—	RMSE, R^2 , Spearman's ρ
Thuy et al. (2024)	2024	Manual	✓	—	—	RMSE
Current	2025	Manual	✓	✓	✓	Balanced DRPS

4.2.1 Output Types

Early work on discrete-level QDE employed classification models, such as support vector machines and Bayesian NNs (Fang et al., 2019; Hsu et al., 2018; Yang & Suyong, 2018). Subsequently, Lin et al. (2019) proposed a discretized regression approach using an LSTM-based NN, where difficulty was predicted as a continuous value between 0 and 1 and mapped to discrete intervals (e.g., [0.0;0.2), [0.2;0.4), etc.).

Zhou and Tao (2020) introduced a multi-task BERT-based model that leverages shared representations across datasets, using a classification head to predict difficulty levels. To reduce the reliance on large labeled datasets in supervised methods, Loginova et al. (2021) proposed an unsupervised QDE method. They leverage the uncertainty in pre-trained question-answering models as a proxy for human-perceived difficulty, computed as the variance over the predictions from an ensemble of classification models.

Benedetto (2023) conducted a benchmarking study comparing traditional machine learning methods and end-to-end NNs across datasets with both discrete and continuous difficulty labels. Their results show that fine-tuned Transformer-based models such as BERT and DistilBERT consistently outperform classical methods. However, these models exclusively employed a discretized regression approach to address the ordinal nature of the labels.

More recently, Thuy et al. (2024) investigated active learning for QDE, demonstrating that comparable performance to fully supervised models can be achieved by labeling only a small fraction of the training data. Yet again, only a discretized regression modeling strategy was considered. As summarized in Table 4.1, prior work has not directly compared discretized regression and classification approaches, and specialized ordinal regression methods remain entirely unexplored.

4.2.2 Metrics

Accuracy is the most widely used evaluation metric for discrete-level QDE, particularly in studies employing classification models, as it aligns with the cross-entropy loss typically used during training. Lin et al. (2019) is the only study using a discretized regression approach that also reports accuracy. However, accuracy fails to account for the ordinal structure of difficulty levels: all misclassifications are treated equally, regardless of their distance from the true label. Even if a prediction is incorrect, it should still be as close as possible to the true difficulty level. Additionally, accuracy is a threshold-based metric, relying solely on the final predicted label rather than the full output distribution—an issue that also applies to metrics such

as the F_1 -score. To partially address this, Hsu et al. (2018) and Lin et al. (2019) report adjacent accuracy, defined as the proportion of predictions within k levels of the true label. However, the choice of k is often arbitrary and dataset-dependent, complicating comparisons across studies.

The most recent works, which adopt discretized regression approaches, report RMSE as their primary evaluation metric, treating difficulty levels as integers. While RMSE reflects the ordinal structure to some extent, it assumes uniform spacing between levels—a condition rarely met in real-world data. For example, in primary school, there is a non-linear increase in difficulty over the years (Coe et al., 2008). This limitation also applies to metrics such as R^2 and Spearman’s rank correlation. Furthermore, since RMSE is used as the loss function in discretized regression models, these models benefit from being evaluated on the same objective they were optimized for, giving them an unfair advantage and introducing a potential bias in comparative evaluations.

In addition to the ordinality aspect, these commonly used metrics fail to account for class imbalance, which is a prevalent issue in discrete-level QDE. In many datasets, mid-range difficulty levels tend to dominate, while questions at the extremes—those that are very easy or very difficult—are underrepresented (Clark et al., 2018; Liang et al., 2019). Standard metrics, which compute aggregate scores across all samples, inherently place greater weight on the majority classes. As a result, model performance on minority classes is often underrepresented, leading to inflated metrics that do not accurately reflect model effectiveness across the full difficulty spectrum. This is particularly problematic in educational contexts, where balanced performance across all difficulty levels is essential to ensure adequate personalized learning experiences.

As a result, the literature ignores the ordinal aspect in the modeling approaches and employs suboptimal evaluation metrics for discrete-level QDE. There is a clear need for an evaluation metric that simultaneously accounts for ordinality and class imbalance, thereby enabling fair comparison across different modeling strategies.

4.3 Balanced Discrete Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) is the most widely adopted scoring rule for evaluating probabilistic forecasts of real-valued variables, such as in precipitation forecasting (Gneiting & Raftery, 2007). It is defined as the integral of the squared difference (i.e., Brier score) between the cumulative distribution function (CDF) of a probabilistic forecast F and the CDF of the observed outcome, at all real-valued thresholds. The

observed outcome is represented as a degenerate distribution, as its CDF is a step function. Formally, given a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$, the CRPS is computed as:

$$\text{CRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{\infty} (F(\hat{y}_i) - \mathbb{1}\{\hat{y}_i \geq y_i\})^2 d\hat{y}_i,$$

where $F(\hat{y}_i)$ denotes the CDF of the forecast and $\mathbb{1}\{\cdot\}$ is the step function.

The CRPS is distance-sensitive, meaning it rewards forecasts that assign higher probability mass to values near the true outcome. Specifically, when the forecast distribution concentrates probability density around the true value, the squared error between the forecast CDF and the observed step function is smaller across the integration range, resulting in a lower (better) score. In other words, even if the predicted value does not exactly coincide with the true outcome, placing substantial probability on neighboring values results in a better score than a prediction that is entirely off-target.

This property makes the CRPS particularly well-suited for ordinal prediction tasks. In such cases, the DRPS serves as a natural extension of the CRPS for discrete outcomes across K ordered categories. The DRPS has only been applied in meteorology (Weigel et al., 2007) and has received little attention in the general field of ordinal regression. It is defined as:

$$\text{DRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K-1} (F_k(\hat{y}_i) - \mathbb{1}\{k \geq y_i\})^2,$$

where $F_k(\hat{y}_i)$ denotes the predicted cumulative probability up to class k . The step function $\mathbb{1}\{\cdot\}$ moves from 0.0 to 1.0 at the position of the ground truth label.

Unlike other evaluation metrics in related work, the DRPS operates on full probability distributions rather than point estimates, enabling a more nuanced assessment of ordinal predictions. Such a probability distribution over the levels is available for all classification and ordinal regression models, but not for the discretized regression model as it only outputs a predicted difficulty level. In the case of deterministic predictions, the output is treated as a degenerate distribution—analogueous to the representation of the observed outcome. Figures 4.1 and 4.2 illustrate how the DRPS is computed for single observations with probabilistic and deterministic model outputs.

Crucially, the DRPS respects the ordinal structure of the prediction task without assuming equal inter-class distances, a notable limitation of metrics

ORDINALITY IN DISCRETE-LEVEL QUESTION DIFFICULTY ESTIMATION: INTRODUCING BALANCED DRPS AND ORDERED LOGITNN

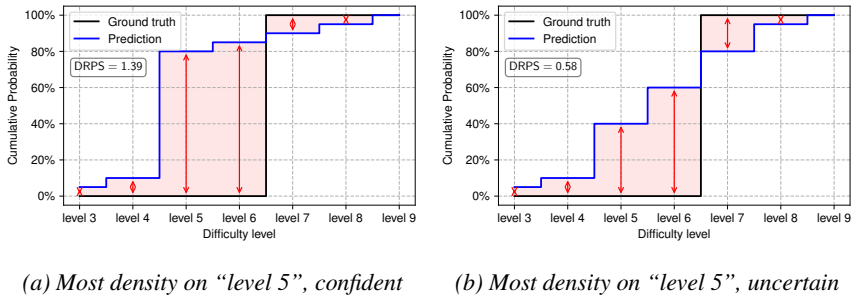


Figure 4.1: Example DRPS calculation with predicted probabilities. The ground truth for this observation is level 7 and both predictions have the most density on the incorrect level 5. (a) assigns high probability to level 5 while (b) is more uncertain and has more density on the neighboring levels, resulting in a better score.

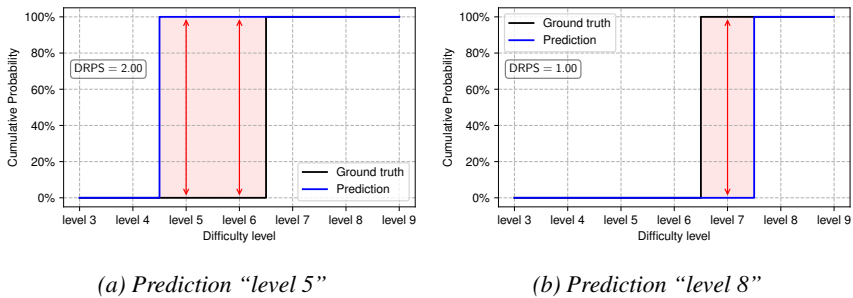


Figure 4.2: Example DRPS calculation with a predicted label. The ground truth for this observation is level 7; (a) predicts level 5 while (b) predicts level 8.

such as RMSE. Additionally, when applied to deterministic predictions, the DRPS reduces to the mean absolute error, providing a direct way to compare deterministic and probabilistic predictions within a unified evaluation metric.

In this work, we introduce the balanced DRPS to address the class imbalance in discrete-level QDE datasets, where extreme difficulty levels are typically underrepresented compared to mid-range levels (Clark et al., 2018; Liang et al., 2019). The popular metrics (accuracy and RMSE) and standard DRPS compute unweighted averages, which overemphasize performance on majority classes and can produce misleadingly high scores on imbalanced data. However, for educational practitioners, robust performance across the full spectrum of difficulty levels—including the rarest—is essential. To ensure fair evaluation, the balanced DRPS weights each ob-

ervation inversely proportional to the prevalence of its true class, $w_i = \frac{1}{\sum_{j=1}^N \mathbb{1}\{y_j=y_i\}}$, thereby giving equal importance to all difficulty levels. Formally:

$$\text{Balanced DRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K-1} w_i (F_k(\hat{y}_i) - \mathbb{1}\{k \geq y_i\})^2.$$

Thus for balanced datasets, the balanced DRPS is equivalent to the standard DRPS.

In conclusion, the balanced DRPS offers a robust and fair evaluation metric for discrete-level QDE by accounting for both ordinal structure and class imbalance. It supports both deterministic and probabilistic predictions and remains neutral to training objectives, making it well-suited for benchmarking across diverse modeling approaches.

4.4 Ordered Logit for NNs

OrderedLogitNN extends the classical ordered logit model to NNs, effectively bridging the gap between econometrics and deep learning. At its core, it is a latent variable model, where each observation is associated with an unobserved continuous utility value y_i^* (Greene & Hensher, 2010) modeled as: $y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$. The observed ordinal outcome y_i is derived from y_i^* through a censoring mechanism, whereby the continuous latent variable is mapped to one of K discrete categories based on a sequence of $K + 1$ increasing threshold values $\{\mu_{-1}, \mu_0, \mu_1, \dots, \mu_{K-1}\}$.

To identify the model parameters, several normalizations are required. First, the thresholds must be increasing $\mu_k > \mu_{k-1}$ to ensure valid (i.e., positive) probabilities. Second, the endpoints of the support are fixed as $\mu_{-1} = -\infty$ and $\mu_{K-1} = +\infty$, covering the entire real line. Third, the error term ϵ_i is assumed to follow a standardized logistic distribution (mean zero, variance $\frac{\pi^2}{3}$). The logistic distribution is preferred over the Gaussian (i.e., probit) for computational convenience, as the derivative has a closed form solution and is readily available as the sigmoid function. Finally, since \mathbf{x}_i includes a bias term, the threshold $\mu_0 = 0$.

The model defines the class probabilities as:

$$P(y_i = k | \mathbf{x}_i) = F(\mu_k - \mathbf{x}_i\boldsymbol{\beta}) - F(\mu_{k-1} - \mathbf{x}_i\boldsymbol{\beta}),$$

with F the cdf of the logistic distribution. An example for $K = 3$ is shown in Figure 4.3.

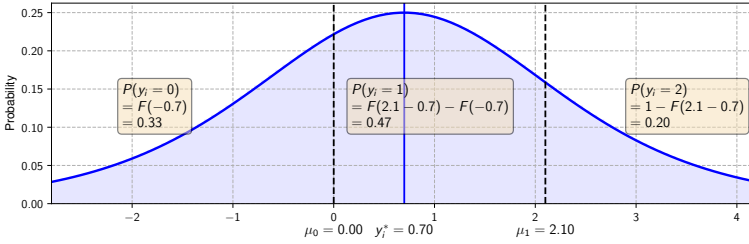


Figure 4.3: *Ordered logit example for three ordinal categories.*

The NN is trained by minimizing the negative log-likelihood (NLL):

$$\text{NLL} = \sum_{i=1}^N \sum_{j=0}^{K-1} m_{ij} \log [F(\mu_k - \mathbf{x}_i \boldsymbol{\beta}) - F(\mu_{k-1} - \mathbf{x}_i \boldsymbol{\beta})],$$

where $m_{ij} = 1$ if $y_i = k$ and 0 otherwise. The thresholds are reparameterized to ensure monotonically increasing values: $\mu_k = \mu_{k-1} + \exp(\delta_k) = \sum_{m=1}^k \exp(\delta_m)$.

The δ_k parameters are initialized such that the ordinal levels have equal probability mass under the logistic distribution, with the first threshold set to zero. The bias term is initialized to lie at the center of this distribution, while all other weights follow standard initialization practices in PyTorch. To facilitate convergence, the learning rates for the δ_k values and the bias term are scaled to be 100 times larger than those of the remaining network parameters.

Importantly, OrderedLogitNN is architecture-agnostic and can be integrated into any NN. Additionally, it makes no assumptions about the distances between ordinal levels, allowing it to flexibly model a wide range of ordered regression problems.

4.5 Existing Approaches for Ordinal Regression with NNs

This section describes existing approaches to handle ordinal regression using NNs. Depending on the specific approach, the amount of ordinal information used and the underlying assumptions vary. This study uses three existing specialized ordinal regression methods that follow the extended binary classification framework, most widely used in the ordinal regression literature (Shi et al., 2023). Note that the methods discussed below are not tied to any specific architecture and can be utilized with any NN.

Let $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be the training dataset consisting of N training examples. Here, $\mathbf{x}_i \in \mathcal{X}$ denotes the i^{th} training example and y_i the corresponding rank, where $y_i \in \mathcal{Y} = \{r_1, r_2, \dots, r_K\}$ with ordered rank $r_K \succ r_{K-1} \succ \dots \succ r_1$. The objective is to find a model that maps $\mathcal{X} \rightarrow \mathcal{Y}$. For example, the ARC dataset has $K = 7$ difficulty levels with an output space $\mathcal{Y} = \{\text{“grade 3”}, \text{“grade 4”}, \dots, \text{“grade 9”}\}$.

4.5.1 Discretized Regression

In the regression approach, also referred to as discretized regression, the K rank indices are treated as numerical values to utilize the ordinal information (see Table 4.1 for references). The model f minimizes the mean squared error loss and predicts a real-valued quantity $f(\mathbf{x}_i) \in \mathbb{R}$ representing a continuous rank estimate, which is then converted to the closest rank index. For example, a regression estimate of 2.7 is converted to index 3 while estimate 5.2 is converted to index 5.

Using a discretized regression approach in an ordinal QDE problem assumes that the inter-level distances are equal. However, this condition is only rarely satisfied in practice (Coe et al., 2008). On the ARC dataset with levels “grade 3” to “grade 9”, for example, such an approach assumes that the jump in difficulty among all grades is identical.

4.5.2 Classification

In the classification approach, the model’s output space is a set of K unordered labels, one for each rank (see Table 4.1 for references). The model is trained to minimize the cross-entropy loss and the predicted rank label is the class with the highest predicted probability. As such, the predicted rank label is $\hat{y}_i = \arg \max_{y_i \in \mathcal{Y}} p(y_i | \mathbf{x}_i)$.

This approach essentially assumes that the difficulty levels are completely independent, hence discarding the available ordinal information. For example, for a question in the ARC dataset with true level “grade 3”, predicting levels “grade 4” and “grade 5” incurs the same loss even though the difference between “grade 3” and “grade 5” is larger than the that between level “grade 3” and “grade 4”.

4.5.3 Ordinal: OR-NN

A popular general machine learning approach to ordinal regression is to cast it as an extended binary classification problem (Li & Lin, 2006), leveraging the relative order among the labels. That is, the ordinal regression task with K ranks is represented as a series of $K - 1$ simpler binary classification

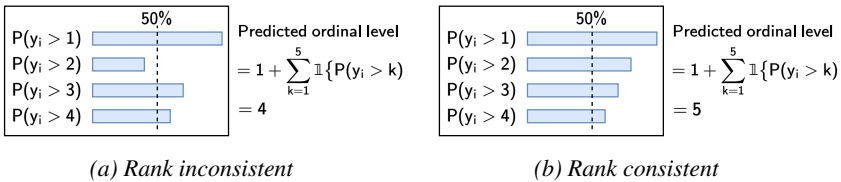


Figure 4.4: **Example of a rank inconsistent and rank consistent prediction.** Both tasks have $K = 5$ levels. Figure adapted from Shi et al. (2023).

sub-problems. For each rank index $k \in 1, 2, \dots, K - 1$, a binary classifier is trained according to whether the rank of a sample is larger than k . As such, all $K - 1$ tasks share the same intermediate layers but are assigned distinct weight parameters in the output layer. In summary, this framework relies on three steps: (i) extending rank labels to binary vectors, (ii) training binary classifiers on the extended labels, and (iii) computing the predicted rank label from the binary classifiers.

In 2016, the authors of Niu et al. (2016) adapted this framework to train NNs for ordinal regression; we refer to this method as OR-NN. More formally, a rank label y_i is first extended into $K - 1$ binary vectors $y_i^{(1)}, \dots, y_i^{(K-1)}$ such that the $y_i^{(k)} \in \{0, 1\}$ indicates whether y_i exceeds rank r_k , for instance, $y_i^{(k)} = \mathbb{1}\{y_i > r_k\}$. Using the extended binary labels, a single NN is trained with $K - 1$ binary classifiers in the output layer to minimize the cross-entropy loss. Based on the binary task predictions, the predicted rank label is $\hat{y}_i = r_{q_i}$. The rank index q_i is given by

$$q_i = 1 + \sum_{k=1}^{K-1} \mathbb{1}\{P(y_i > r_k) > 0.5\},$$

where $P(y_i > r_k) \in [0, 1]$ is the predicted probability of the k th binary classifiers in the output layer.

However, the authors pointed out that OR-NN can suffer from rank inconsistencies among the binary tasks such that the predictions for individual binary tasks may disagree. For example, on the RACE++ dataset, it would be contradictory if the first binary task predicts that the difficulty is not higher than middle school level while the second binary task predicts it to be more difficult than high school level. This inconsistency could lead to suboptimal results when combining the $K - 1$ predictions to obtain the estimated difficulty level. Figure 4.4 provides an example of a rank consistent and inconsistent prediction. In response, two methods have been proposed that overcome this drawback of rank inconsistency: CORAL and CORN.

4.5.4 Ordinal: CORAL and CORN

CORAL (Cao et al., 2020) achieves rank consistency by imposing a weight-sharing constraint in the last layer. Instead of learning distinct weights between each unit in the penultimate layer and each output unit, CORAL enforces that the $K - 1$ binary tasks share the same weight parameters to the units in the penultimate layer. In addition, CORAL learns independent bias terms for each output unit as opposed to a single bias term for the output layer.

CORAL uses a cross-entropy loss over the $K - 1$ binary classifiers and the authors show theoretically that by minimizing this loss function, the learned bias terms of the output layer are non-increasing such that $b_1 \geq b_2 \geq \dots \geq b_{K-1}$. Consequently, the predicted probabilities of the $K - 1$ tasks are decreasing which ensures that the output reflects the ordinal information and is rank consistent. All other steps are identical to the extended binary classification framework.

However, while CORAL outperforms the OR-NN method in age prediction (Cao et al., 2020), the weight-sharing constraint may restrict the expressiveness and capacity of the NN.

More recently, the authors of Shi et al. (2023) proposed CORN, which guarantees rank consistency without restricting the NN's expressiveness and capacity. CORN achieves rank consistency by a novel training scheme which uses conditional training sets in order to obtain the unconditional rank probabilities.

More formally, CORN constructs conditional training subsets such that the output of the k th binary task $f_k(\mathbf{x}_i)$ represents the conditional probability $f_k(\mathbf{x}_i) = P(y_i > r_k \mid y_i > r_{k-1})$. For $k \geq 2$, the conditional subsets consist of observations where $y_i > r_{k-1}$. When $k = 1$, $f_1(\mathbf{x}_i)$ represents the initial unconditional probability $P(y_i > r_1)$ based on the complete dataset. The transformed unconditional probabilities can then be computed by applying the chain rule for probabilities: $P(y_i > r_k) = \prod_{j=1}^k f_j(\mathbf{x}_i)$.

Since $\forall j, 0 \leq f_j(\mathbf{x}_i) \leq 1$, we have $P(y_i > r_1) \geq P(y_i > r_2) \geq \dots \geq P(y_i > r_{K-1})$, which guarantees rank consistency among the $K - 1$ binary tasks. During model training, CORN minimizes the cross-entropy loss over the binary tasks. All other steps are identical to the extended binary classification framework.

4.6 Experiments

4.6.1 Data

We evaluate our models on two multiple-choice question (MCQ) datasets: RACE++ and ARC. These datasets vary in domain and granularity of difficulty levels.

RACE++ (Lai et al., 2017; Liang et al., 2019) is a dataset of reading comprehension MCQs. Questions are labeled with one of three difficulty levels—1 (middle school), 2 (high school), and 3 (university level)—which we treat as ground truth labels for QDE. The label distribution is imbalanced: 25% of the questions are labeled as level 1, 62% as level 2, and 13% as level 3. The dataset is partitioned into training, validation, and test sets containing 100,568; 5599; and 5642 questions, respectively.

ARC (Clark et al., 2018) is a dataset of science MCQs across grades 3 through 9. Question difficulty is indicated by the target grade level (i.e., 7 levels), which we use as ground truth. The training, validation, and test splits contain 3358, 862, and 3530 questions, respectively. The distribution is highly imbalanced: level 8 appears approximately 1400 times, level 5 about 700 times, and level 6 only 100 times. To decrease this imbalance, we follow Benedetto (2023) and downsample the two most frequent levels to 500 examples each, resulting in a partially balanced training set of 2293 questions.

4.6.2 Model Architecture

We focus on end-to-end Transformer-based NNs, as they have been shown to outperform traditional NLP approaches that rely on separate feature engineering and modeling stages (Benedetto, 2023). We fine-tune the Transformer BERT (“bert-base-uncased”) on the task of QDE, stacking an output layer on top of the pre-trained language model. During fine-tuning, both the weights of the output head and the pre-trained model are updated. We follow the input encoding of Benedetto (2023) and concatenate the question and the text of all the possible answer choices in a single sentence, divided by separator tokens.

Additionally, we investigate the performance of two baselines which serve as a lower bound on performance: (i) Random and (ii) Majority. The Random baseline randomly predicts a difficulty level, while the Majority baseline consistently predicts the majority level in the training set.

The experiments are implemented in PyTorch (Ansel et al., 2024) using the HuggingFace (Wolf et al., 2020) package, and results are averaged over five independent runs with random seeds.

Table 4.2: Results on RACE++ (3 levels).

Output type	Bal. DRPS ↓	Bal. DRPS ↓ (degenerate)	RMSE ↓	Accuracy ↑
Random	0.893 ± 0.009	0.893 ± 0.009	1.024 ± 0.005	0.334 ± 0.003
Majority	0.667 ± 0.000	0.667 ± 0.000	0.616 ± 0.000	0.620 ± 0.000
Regression	0.167 ± 0.003	0.167 ± 0.003	0.391 ± 0.004	0.853 ± 0.003
Classification	0.133 ± 0.003	0.170 ± 0.004	0.402 ± 0.006	0.847 ± 0.004
OR-NN	0.131 ± 0.004	0.168 ± 0.004	0.400 ± 0.008	0.847 ± 0.006
CORAL	0.201 ± 0.003	0.185 ± 0.004	0.476 ± 0.018	0.782 ± 0.017
CORN	0.127 ± 0.003	0.164 ± 0.001	0.397 ± 0.004	0.851 ± 0.003
OrderedLogitNN	0.130 ± 0.001	0.162 ± 0.002	0.384 ± 0.005	0.861 ± 0.003

4.7 Results and Discussion

4.7.1 Balanced DRPS

Tables 4.2 and 4.3 present the results for the RACE++ and ARC datasets, which contain 3 and 7 difficulty levels, respectively. Balanced DRPS with probabilistic inputs is the main evaluation metric, as these predictions express uncertainty and are particularly valuable in downstream decision-making. In addition, we report the balanced DRPS computed using degenerate distributions—i.e., where all probability mass is placed entirely on the predicted level. This setup removes any representation of uncertainty from the predictions, thereby altering the scores for both classification and ordinal regression methods. Notably, the baselines and the discretized regression model remain unaffected in this setting, as they do not express uncertainty. Recall that for both metrics, lower is better. Furthermore, we include the commonly used but flawed metrics RMSE and accuracy.

On RACE++ with 3 levels (Table 4.2), the classification model performs comparably to the ordinal methods OR-NN, CORN, and OrderedLogitNN in terms of balanced DRPS. The discretized regression model, by contrast, shows slightly inferior performance. Interestingly, the CORAL model underperforms significantly, likely due to its weight-sharing constraint limiting the NN’s capacity. Nonetheless, all models substantially outperform the baseline approaches. We hypothesize that the small differences in performance across models are due to the limited number of ordinal levels.

The ARC dataset with 7 levels (Table 4.3) presents a more challenging setting, as it includes a greater number of difficulty levels and exhibits more pronounced class imbalance. Here, the OrderedLogitNN model con-

Table 4.3: *Results on ARC (7 levels).*

Output type	Bal. DRPS ↓	Bal. DRPS ↓ (degenerate)	RMSE ↓	Accuracy ↑
Random	2.296 ± 0.017	2.296 ± 0.017	2.730 ± 0.006	0.144 ± 0.002
Majority	2.286 ± 0.000	2.286 ± 0.000	2.195 ± 0.000	0.409 ± 0.000
Regression	1.030 ± 0.015	1.030 ± 0.015	1.412 ± 0.013	0.356 ± 0.003
Classification	0.720 ± 0.003	1.038 ± 0.005	1.522 ± 0.013	0.421 ± 0.005
OR-NN	0.722 ± 0.005	1.005 ± 0.003	1.435 ± 0.009	0.400 ± 0.002
CORAL	0.963 ± 0.003	1.440 ± 0.020	2.083 ± 0.033	0.156 ± 0.005
CORN	0.725 ± 0.007	1.046 ± 0.012	1.428 ± 0.004	0.389 ± 0.008
OrderedLogitNN	0.674 ± 0.004	0.980 ± 0.007	1.468 ± 0.007	0.393 ± 0.006

siderably outperforms all other methods. For the remaining methods, the insights are consistent with those observed on RACE++.

When restricting the predictions to degenerate distributions, scores generally deteriorate (see the third column in Tables 4.2 and 4.3) because balanced DRPS is designed to reward well-calibrated probabilistic predictions while penalizing overconfident, incorrect ones. This shift brings the regression model’s performance in line with the classification model, OR-NN, and CORN. OrderedLogitNN again performs on par for RACE++ and performs substantially better on ARC.

When considering RMSE and accuracy—metrics that are poorly suited for ordinal prediction tasks (see Section 4.2.2)—we observe that models closely related to these objectives unsurprisingly achieve the best performance. On the RACE++ dataset, the results are close and OrderedLogitNN performs comparably to or even slightly better than the regression and classification models. In contrast, on ARC, the regression approach achieves the lowest RMSE, while classification achieves the highest accuracy. These results are expected, as the regression method is directly optimized with RMSE loss while the classification method entirely omits the ordinal information, just like the accuracy metric.

4.7.2 Confusion Matrix

To further investigate the behavior of the models, Figure 4.5 presents confusion matrices for the ARC dataset, the more complex task in this study. These matrices are normalized by the true class (i.e., row-wise) and are based on discrete predicted levels—rather than full probability distributions—thus aligning with the evaluation setting of the balanced

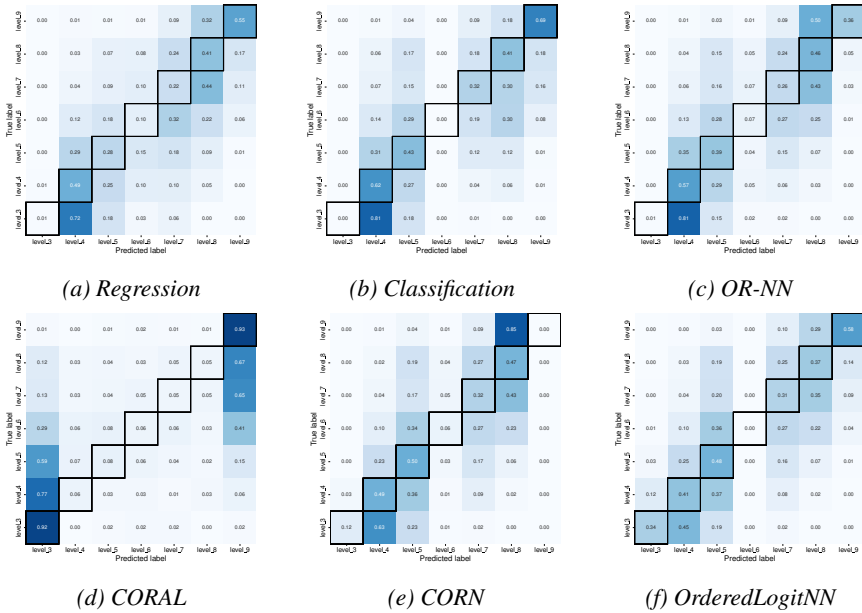


Figure 4.5: *Confusion matrices on the ARC dataset. The results are shown for all models, normalized over the true labels (rows).*

DRPS with degenerate predictions.

The results indicate the inherent difficulty of the task, as demonstrated by the substantial deviations from the diagonal (highlighted in black). All models hardly predict level 6 as these observations are least present in the training set. Notably, the CORAL model exhibits highly atypical behavior, predicting exclusively at the extreme levels (3 and 9). This pattern suggests that the model fails to converge to a meaningful solution, consistent with its poor performance on the balanced DRPS metric.

Among the remaining models, all except OrderedLogitNN show concentrated errors in specific off-diagonal cells, i.e., predicting level 4 instead of level 3, or level 8 instead of level 9. These errors are associated with the outermost classes (levels 3 and 9), which are often neglected entirely by the models. In contrast, OrderedLogitNN stands out as the only method that successfully captures both extremes of the ordinal scale, avoiding these consistent misclassifications.

4.8 Conclusion

This study approaches discrete-level QDE through the lens of ordinal regression, reflecting the inherent ordering of difficulty levels from easiest to hardest. Prior work in this area has ignored this ordinal structure, both in the choice of modeling paradigms and in the design of evaluation metrics.

To address these gaps, we benchmark three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced DRPS, a novel metric that captures both ordinality and class imbalance. Moreover, we propose a novel ordinal regression model OrderedLogitNN, extending the ordered logit model from econometrics to NNs. We fine-tune the Transformer model BERT on the RACE++ and ARC datasets.

Experimental results indicate that OrderedLogitNN considerably outperforms existing methods on more complex tasks while performing comparably on simpler ones, both on probabilistic and degenerate predictions. Probabilistic predictions express uncertainty and are particularly valuable in downstream human-in-the-loop tasks such as selective prediction or active learning (Thuy & Benoit, 2024; Thuy et al., 2024). These findings underscore OrderedLogitNN as a highly attractive approach for discrete-level QDE. The learning dynamics of OrderedLogitNN deserve further investigation and future work can assess the impact of adapting the weight initializations and learning rate multiplier.

Our work has important practical implications for QDE, as more reliable difficulty estimation enables scalable personalized learning paths in educational platforms. Furthermore, it has broader relevance for the automated evaluation of assessment content with ordinal labels, such as essay correction. The balanced DRPS provides a principled foundation for evaluating such systems in production and for future research. Finally, OrderedLogitNN’s robustness makes it well-suited for integration in educational applications where strong performance over the entire label range is critical.

Reflections and Clarifications

Kullback-Leibler Divergence

In this ordinal regression setting, we compare the ground truth difficulty level (i.e., a degenerate distribution) to the prediction, which may either be a single difficulty level (also a degenerate distribution) or a probability distribution over the difficulty levels. The proposed balanced DRPS provides a meaningful way to quantify the discrepancy between these two distribu-

tions.

It is important to note that using the Kullback-Leibler (KL) divergence is not suitable for this application. The KL divergence is defined as

$$D_{\text{KL}}(P \parallel Q) = \sum_{i=1}^C P(i) \log \frac{P(i)}{Q(i)},$$

where $P(i)$ represents the probability for class i under distribution P ; analogously for $Q(i)$. It measures the information gained when moving from a reference distribution (Q) to another distribution (P).

However, KL divergence has two major limitations in this context. First, although it quantifies dissimilarity between distributions, it is usually not considered a true metric because it is not symmetric. The KL divergence from P to Q is not the same as from Q to P . Second, when one of the distributions is degenerate, the KL divergence becomes infinite unless the two distributions are identical. If P is a degenerate distribution, there is a class i for which $P(i) = 0$. If Q is not a degenerate distribution and $Q(i) > 0$, the term $\log 0$ arises, which is undefined. To make the expression valid, $Q(i)$ would also have to be zero, an unrealistic condition in most cases. Therefore, the KL divergence fails to exhibit the desired properties of a metric for this setting.

Neural Network Settings

Hyperparameters have been tuned individually for each NN architecture using Optuna, with the objective of minimizing the balanced DRPS on the validation set. We tune the learning rate (range [1e-5, 5e-4]; log-scale) and weight decay (range [1e-3, 5e-1]; log-scale) for 20 trials using the default sampler (Tree-structured Parzen Estimator). Moreover, we set the batch size to 128, maximum sequence length to 512, warmup ratio to 0.1, and we train for 3 epochs (including the pre-trained base) with early stopping.

For the OrderedLogitNN model, the threshold parameters δ_k were also optimized via stochastic gradient descent. In preliminary experiments, we observed that these thresholds converged very slowly. Since each overall threshold is defined as the cumulative sum of the δ_k parameters, all parameters must move coherently in the same direction and with sufficient magnitude for the threshold to meaningfully adjust, likely causing the slow convergence. To address this, we increased the learning rate magnitude for the δ_k parameters. After testing values of 1e1, 1e2, and 1e3, a multiplier of 1e2 yielded the most effective results and was therefore used in all subsequent experiments. Nevertheless, this value should ideally be tuned as an

additional hyperparameter, suggesting that the method's performance could be further enhanced with more extensive optimization.

Acknowledgments

This study was supported by the Research Foundation Flanders (FWO) (grant number 1S97022N).

References

- Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., . . . Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. <https://doi.org/10.1145/3620665.3640366>
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series, 2014(2)*, 1–8. <https://doi.org/10.1002/ets2.12042>
- Benedetto, L. (2023). A quantitative study of nlp approaches to question difficulty estimation. *International Conference on Artificial Intelligence in Education*, 428–434. https://doi.org/10.1007/978-3-031-36336-8_67
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1–37. <https://doi.org/10.1145/3556538>
- Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140, 325–331. <https://doi.org/10.1016/j.patrec.2020.11.008>
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*. <https://doi.org/10.48550/arXiv.1803.05457>

- Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). Relative difficulty of examinations in different subjects. *Durham: CEM centre*.
- Fang, J., Zhao, W., & Jia, D. (2019). Exercise difficulty prediction in online education systems. *2019 International Conference on Data Mining Workshops (ICDMW)*, 311–317.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477), 359–378.
- Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.
- Hambleton, R. K. (1991). *Fundamentals of item response theory*. Sage.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, 12(3), 38–47.
- Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6), 969–984.
- Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017, September). RACE: Large-scale ReAding comprehension dataset from examinations. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D17-1082>
- Lane, S., Raymond, M. R., Haladyna, T. M., et al. (2016). *Handbook of test development* (Vol. 2). Routledge New York, NY.
- Li, L., & Lin, H.-T. (2006). Ordinal regression by extended binary classification. *Advances in neural information processing systems*, 19.
- Liang, Y., Li, J., & Yin, J. (2019). A new multi-choice reading comprehension dataset for curriculum learning. *Asian Conference on Machine Learning*, 742–757.
- Lin, L.-H., Chang, T.-H., & Hsu, F.-Y. (2019). Automated prediction of item difficulty in reading comprehension using long short-term memory. *2019 International Conference on Asian Language Processing (IALP)*, 132–135.
- Loginova, E., Benedetto, L., Benoit, D., & Cremonesi, P. (2021). Towards the application of calibrated transformers to the unsupervised es-

- timation of question difficulty from text. *RANLP 2021*, 846–855. https://doi.org/https://doi.org/10.26615/978-954-452-072-4_097
- Niu, Z., Zhou, M., Wang, L., Gao, X., & Hua, G. (2016). Ordinal regression with multiple output cnn for age estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4920–4928.
- Shi, X., Cao, W., & Raschka, S. (2023). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3), 941–955. <https://doi.org/10.1007/s10044-023-01181-9>
- Thuy, A., & Benoit, D. F. (2024). Explainability through uncertainty: Trustworthy decision-making with neural networks. *European Journal of Operational Research*, 317(2), 330–340. <https://doi.org/10.1016/j.ejor.2023.09.009>
- Thuy, A., Loginova, E., & Benoit, D. F. (2024). Active learning to guide labeling efforts for question difficulty estimation. *arXiv preprint arXiv:2409.09258*. <https://doi.org/10.48550/arXiv.2409.09258>
- Thuy, A., Loginova, E., & Benoit, D. F. (2025). Ordinality in discrete-level question difficulty estimation: Introducing balanced drps and orderedlogitnn. *Second Workshop on Automated Evaluation of Learning and Assessment Content, Vol. 4006*. <https://ceur-ws.org/Vol-4006/paper4.pdf>
- Van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Springer. <https://doi.org/10.1007/0-306-47531-6>
- Wang, Q., Liu, J., Wang, B., & Guo, L. (2014). A regularized competition model for question difficulty estimation in community question answering services. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1115–1126. <https://doi.org/10.3115/v1/D14-1118>
- Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135(1), 118–124. <https://doi.org/10.1175/MWR3280.1>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

- Yang, H., & Suyong, E. (2018). Feature analysis on english word difficulty by gaussian mixture model. *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, 191–194.
- Zhou, Y., & Tao, C. (2020). Multi-task bert for problem difficulty prediction. *2020 international conference on communications, information system and computer engineering (cisce)*, 213–216. <https://doi.org/10.1109/CISCE50729.2020.00048>

ORDINALITY IN DISCRETE-LEVEL QUESTION DIFFICULTY ESTIMATION:
INTRODUCING BALANCED DRPS AND ORDERED LOGIT NN

5

Leveraging Misconceptions with In-Context Learning to Simulate Students with Role-Playing LLMs

This chapter presents ongoing work, which was later revised and submitted to the International Conference on Language Resources and Evaluation (LREC). The manuscript has been co-authored with Dr. Luca Benedetto (University of Cambridge, Institut Polytechnique de Paris), Dr. Ekaterina Loginova (Dedalus Healthcare), and Prof. dr. Dries F. Benoit (Ghent University).

Abstract

Large Language Models (LLMs) have recently been investigated to simulate student responses to exam questions. This approach, referred to as *virtual pretesting*, offers an alternative to conventional pretesting, which is costly and time-consuming. Prior studies have focused on zero-shot role-playing, where an LLM is prompted to imitate students of varying levels, but such methods have shown limited effectiveness. This work introduces a framework that enhances LLM-based student simulation through in-context learning, drawing on previous question-answer records to provide the model with richer information about students' skills and miscon-

ceptions. The results indicate that adding contextual information does not consistently improve the LLMs' student replication performance. Furthermore, the optimal settings for building additional context do not generalize over LLMs, thus stressing the need for prompt engineering each model. While role-playing produces simulated responses with strong Monotonicity scores, these outputs do not yield reliable question difficulty estimates under IRT. Overall, our findings are in line with earlier research: although LLMs show some promise in simulating student behavior, their current capabilities are insufficient for piloting high-stakes educational assessments.

5.1 Introduction

Keeping students engaged with appropriately challenging questions is essential for sustaining motivation and fostering effective learning. Personalized learning systems, such as computerized adaptive testing and online learning platforms, address this challenge by tailoring questions to a student's skill level. A key component of these systems is Question Difficulty Estimation (QDE), which ensures optimal question selection to enhance student engagement and learning outcomes.

Traditionally, QDE has relied on pretesting (Lane et al., 2016), in which new questions are embedded within exams without contributing to students' scores. The responses to these questions, combined with data from other test items, are then used to calibrate question difficulty using statistical models, most commonly Item Response Theory (IRT) (Hambleton, 1991). Although pretesting ensures reliable difficulty estimation, it is time-intensive, costly, and carries the risk of prematurely exposing assessment content.

To mitigate these limitations, previous research has explored supervised machine learning (ML) approaches to QDE (Benedetto, 2023; Benedetto et al., 2023). The current state of the art involves fine-tuning pre-trained encoder-only Transformer models, such as BERT (Devlin et al., 2019), on large collections of calibrated questions. Although this approach leverages transfer learning to generalize difficulty estimation to new items, it still depends on substantial amounts of calibrated training data, which is expensive and difficult to acquire.

Advances in Large Language Models (LLMs) have opened up alternative avenues for QDE by enabling virtual pretesting. In this paradigm, LLMs simulate student responses at various proficiency levels, producing question-answer records that substitute for traditional pretesting data. Two main strategies have been investigated: (i) Multi-LLM simulation: exploiting the natural variation in accuracy across different LLMs, ranging from

low to high-performing models (Park et al., 2024); (ii) Role-playing with a single LLM: prompting a single LLM to generate responses as if it were a student of a specific skill level, ranging from low to high proficiency (Benedetto et al., 2024). In essence, multi-LLM simulation uses the different “skill levels” of various LLMs, while role-playing forces one LLM to answer by simulating different student skill levels.

Prior work merely adopts zero-shot prompting, thereby avoiding reliance on real question–answer records that may be unavailable in practice. While the multi-LLM approach is competitive to supervised fine-tuning on some datasets, its dependence on tens of LLMs makes it computationally expensive and unsuitable for large-scale applications. Conversely, zero-shot role-playing with a single LLM is more efficient but exhibits limited alignment with real student behavior, reducing its reliability for practical use in educational assessment.

To address these challenges, we propose a role-playing framework that enhances virtual pretesting with *in-context learning*. Rather than relying solely on zero-shot prompting, our approach processes real question–answer records automatically collected from learning platforms. These records are used to provide the LLM with rich information about relevant student skills and misconceptions, with the aim of enabling the LLM to more accurately replicate realistic response patterns. Compared with calibrated questions required for supervised fine-tuning, such question–answer records are cheaper and easier to obtain. Importantly, working with this data also allows us to directly evaluate the degree to which LLMs replicate authentic student behavior as an intermediate step, rather than only assessing the final quality of difficulty estimates.

Our contributions are twofold. First, we introduce a framework that integrates contextual information about student understanding into LLM-based role-playing for virtual pretesting. Second, we conduct a systematic benchmark of eight LLMs, encompassing two open-weight model families, on the DBE-KT22 dataset, and compare their performance against zero-shot simulation baselines.

5.2 Related Work

Natural language processing (NLP) has been widely applied to QDE to reduce reliance on manual calibration (Attali et al., 2014) and pretesting (Lane et al., 2016), both of which are costly and time-consuming. The predominant approach is to train supervised models that predict item difficulty directly from question text (AlKhuyaey et al., 2021; Benedetto et al., 2023). Earlier methods rely on traditional ML algorithms and handcrafted

LEVERAGING MISCONCEPTIONS WITH IN-CONTEXT LEARNING TO SIMULATE STUDENTS WITH ROLE-PLAYING LLMs

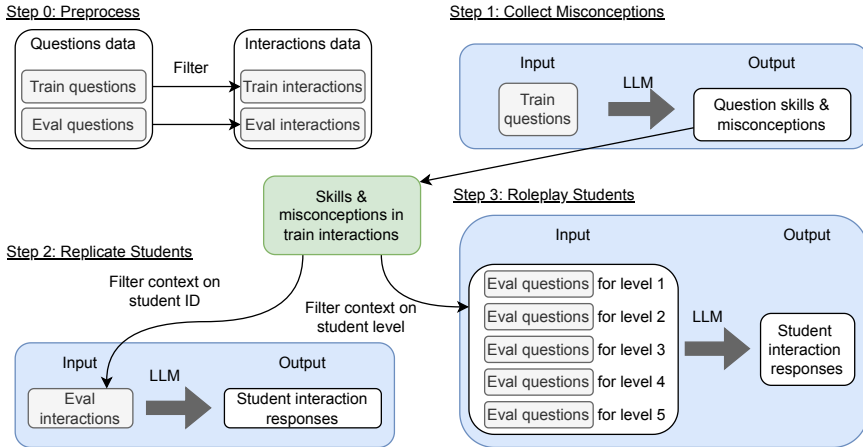


Figure 5.1: Overview of the methodology.

features, including linguistic indicators (Beinborn et al., 2015), word embeddings (Hsu et al., 2018; Yaneva et al., 2019, 2020), and TF-IDF representations (Benedetto et al., 2020). More recent work has established fine-tuning of pre-trained encoder-only Transformer models as the state of the art (Benedetto, 2023; Thuy et al., 2025). Once trained, these models allow for rapid calibration of unseen items, substantially reducing the need for manual calibration and pretesting.

Despite these advances, supervised fine-tuning requires large labeled datasets containing thousands of calibrated items, which are rarely available in practice. Several approaches attempt to alleviate this data bottleneck. For example, Loginova et al. (2021) propose an unsupervised method based on additional pre-training combined with pairwise difficulty estimation, while Thuy et al. (2024) employ active learning to approach supervised-level performance with substantially fewer labeled examples.

More recently, researchers have begun to explore virtual pretesting, in which student responses are simulated and used to estimate item difficulty. This has been pursued either by prompting generative LLMs (Benedetto et al., 2024; Park et al., 2024) or by fine-tuning encoder-based LLMs on existing student response logs (Maeda, 2025; Uto et al., 2025). Within the generative setting, two strategies dominate: Multi-LLM simulation, which leverages the diverse performance levels of different LLMs (Park et al., 2024), and Role-playing with a single LLM, where one model is prompted to imitate students across a range of proficiency levels (Benedetto et al., 2024; Liu et al., 2025; Säuberli et al., 2025).

Studies focusing on role-playing consistently adopt zero-shot prompt-

ing and generally express skepticism regarding its feasibility for high-stakes applications. They emphasize that LLMs, in the absence of contextual information, fail to capture the cognitive mechanisms underlying student responses, and therefore should not be relied upon for piloting educational assessments. Collectively, this line of work highlights the need to move beyond zero-shot prompting and to enrich LLM simulations with student-specific contextual information.

5.3 Methodology

Our methodology evaluates the ability of LLMs to replicate and simulate student learning behavior.

The approach consists of two main stages: Student Behavior Replication (Section 5.3.2) and Student Behavior Roleplay (Section 5.3.3). Both stages leverage in-context learning, supported by a variety of example selection strategies (Section 5.3.4), to condition the model’s predictions with information about relevant skills and misconceptions. The framework is model-agnostic and does not require access to LLMs’ internal parameters. Figure 5.1 illustrates the overall framework.

As a preliminary step, we compile a database of skills and misconceptions associated with each question’s correct answer and distractors (Section 5.3.1). This information is later provided to the LLM in the system prompt to enrich its predictions. Importantly, questions are partitioned into train/validation/test splits, with corresponding splits applied to student interactions. As a result, the same student ID may appear in both training and evaluation sets, allowing us to assess how well the LLMs can simulate individual students in a setting analogous to knowledge tracing (KT) (Shen et al., 2024).

While one could directly proceed from collecting misconceptions to Roleplay, doing so would not reveal which model configurations are most effective at replicating real student behavior. For this reason, we introduce Student Behavior Replication as an intermediate step before moving to Roleplay.

5.3.1 Collecting Skills and Misconceptions

The first step involves constructing a mapping between questions, answer options, and the skills or misconceptions underlying those options. To this end, we employ prompt engineering with a generative LLM. For each question, the model is prompted to identify the skills required to select the correct answer, and list possible misconceptions that could lead a student to

choose each distractor.

The result is a structured dictionary linking each question–answer pair to either a skill (for the correct option) or a misconception (for distractors). The full prompt template is provided in Appendix 5.A.1. For this step, we use the reasoning model o4-mini (checkpoint “o4-mini-2025-04-16”) with a default temperature of 1.0. Note that the resulting skills and misconceptions are not manually evaluated.

5.3.2 Student Behavior Replication

In the replication stage, the LLM is prompted to predict whether *one* specific student will respond correctly to a new question, provided their student level and a selection of skills and misconceptions from their prior interactions as an indication of their current understanding. As such, the LLM predicts on a per-interaction basis. This stage evaluates whether the LLM can accurately reproduce individual student behavior, providing a direct signal of its student-simulation capability.

Although effective replication is critical for virtual pretesting, prior work has not explicitly assessed it, as existing studies ignore previous interactions. Our framework addresses this gap by treating Replication as an intermediate evaluation step that is more resource-efficient than proceeding directly to Roleplay.

Within replication, the LLM adopts one of two personas:

- *Student Persona*: the model is instructed to act as a student at a given proficiency level, reasoning from their skills and misconceptions to select an answer.
- *Teacher Persona*: the model is instructed to act as a teacher, using knowledge of a student’s skills and misconceptions to predict how that student would answer.

Proficiency levels are computed with IRT on training interactions. Prompt templates for both personas are available in Appendix 5.A.2.

Replication shares similarities with KT (Shen et al., 2024), which predicts student performance across a sequence of learning interactions. However, our setting differs from KT as it treats proficiency as static, without modeling temporal learning dynamics. We only select examples previously solved in time because Replication is used as an intermediate step towards Role-playing.

5.3.3 Student Behavior Roleplay

This stage evaluates the LLM’s ability to simulate the behavior of a prototypical student at a certain proficiency level, rather than *one* specific student. The objective is to generate synthetic student–question interactions that reflect realistic performance patterns across different ability levels. These synthetic interactions can then serve as input to downstream psychometric models, such as IRT, for estimating item difficulty.

In this setting, the evaluation set of questions is combined with a pre-defined number of proficiency levels (five in our experiments, though this choice is flexible). Each question is answered once per proficiency level. For example, with five levels and 20 questions, the model produces 100 simulated interactions.

As in replication, the LLM is prompted under both the Student and Teacher personas, with identical prompt templates (Appendix 5.A.2). This alignment ensures that replication provides a strong intermediate signal for identifying configurations that transfer well to Roleplay.

5.3.4 In-context Example Selection Strategies

A central component of our framework is the dynamic construction of contextual information about skills and misconceptions. Each example selection strategy returns a list of k interactions, which are mapped (via the dictionary described in Section 5.3.1) to k skills and misconceptions and inserted into the LLM’s system prompt.

We distinguish between individual-based and group-based selectors: For Individual-Based Selectors (Replication), we use the interaction history of a single student:

- Random: Randomly sample k past interactions from the target student.
- Knowledge Concepts (KC): Identify all previously answered questions and select the k most conceptually similar ones (based on available knowledge concepts) to the target question, returning the corresponding k student interactions.

For Group-Based Selectors (Roleplay), we use the interaction history of a group of students who share the same proficiency level:

- Random: Randomly sample k interactions from the pool of students at the target proficiency level.

- **Knowledge Concepts:** Identify all questions answered by students at that level, select the k most conceptually similar ones (based on available knowledge concepts) to the target question, and return a random interaction for each.

Our methodology fits within the broader paradigm of in-context learning for prompt engineering (Dong et al., 2024), but differs from conventional few-shot prompting. Instead of presenting standard input–output exemplars, the prompts provide contextualized representations of skills and misconceptions relevant to the target interaction.

We deliberately restrict our selection strategies to those that generalize from Replication to Roleplay. For example, a Recency selector—commonly used in KT—would exploit temporal ordering, but this does not meaningfully extend to Roleplay across a group of students. That is, it would not make sense to look at the most recently solved question over an entire group of students, as students are on different learning trajectories.

5.3.5 IRT Estimation

Based on a set of student-question interactions, we apply *theories of testing* to model question difficulty and obtain a numerical estimate. The most widely used framework for this purpose is IRT.

IRT (Hambleton, 1991) models latent traits for both students and questions. In its simplest form, the one-parameter model (i.e., “Rasch Model” (Rasch, 1993)), each student i is assigned a skill level θ_i and each question j is assigned a difficulty level b_j . These latent traits are estimated through a maximum likelihood procedure. At the core of the model lies the item response function, which represents the probability P_{correct} that a student i with skill level θ_i correctly answers the question j with difficulty b_j .

$$P_{\text{correct}} = \frac{1}{1 + e^{-1.7 \cdot (\theta_i - b_j)}}$$

Intuitively, a student with a given skill level θ_i has a lower probability of correctly answering more difficult questions (i.e., those with higher b_j).

More advanced IRT models introduce additional latent traits for questions. The discrimination parameter a determines the steepness of the item response function, while the guess factor c represents the probability that a student correctly answers by guessing. The general item response function can thus be written as:

$$P_{\text{correct}} = c_i + \frac{1 - c_i}{1 + e^{-1.7 \cdot a_i \cdot (\theta_i - b_j)}},$$

which reduces to the one-parameter model when $a_i = 1$ and $c_i = 0$.

5.4 Experimental Setup

5.4.1 Dataset

The DBE-KT22 dataset (Abdelrahman et al., 2022), collected from a relational databases course at the Australian National University, is one of the only publicly available resources that provides both student-question interaction data and the corresponding question text with answer options. The dataset consists of multiple-choice questions (MCQs), for which we compute question difficulty using IRT, regarded as the gold standard for the regression-based QDE task.

To ensure consistency, we restrict the dataset to questions with four answer options and students with at least 30 recorded interactions. For students who answered the same question multiple times, only their first attempt is retained. The dataset is split by question ID into training (92 questions), validation (23 questions), and test (39 questions) sets.

The training interactions, corresponding to the training questions, are used for in-context learning and include 65,494 responses from 988 students. Evaluation sets are subsampled and stratified across five proficiency levels: the small validation set contains 100 interactions, while the large validation and test sets each contain 500 interactions.

5.4.2 Evaluation Metrics

Student Replication is essentially a binary classification task, as the LLM is prompted to predict whether a student will answer correctly. There is a class imbalance as the students correctly answer about 70% of the interactions, while the three distractors represent the remaining 30% of the interactions. Therefore, we consider the balanced accuracy score between the LLM’s response and the actual student’s response.

For Student Roleplay, we employ two metrics, assessing the LLM question responses directly or assessing the question difficulty estimates obtained with IRT. The main metric we use to evaluate QDE with continuous levels, a regression task, is Root Mean Squared Error (RMSE) between the gold standard difficulty and the LLM’s estimates after processing its simulated interactions with IRT. RMSE is the most commonly used metric in the literature when handling continuous levels (Benedetto et al., 2023).

To gain further insights in the LLM Roleplay behavior, we assess the LLM’s simulated interactions. The Monotonicity metric (M) (Benedetto et al., 2024) builds on the idea that the LLM should predict that higher student levels answer more questions correctly than lower student levels. Therefore, it evaluates the correlation $\rho_{L,T}$ between the LLM’s response

correctness per student level $\mathbf{L} = (a_1, a_2, \dots, a_5)$ and those observed in the training set \mathbf{T} , penalizing non-monotonic behavior in the correctness sequence. The penalty for non-monotonicity (P) is calculated as: $\sum_{i=1}^4 \sqrt{|a_{i+1} - a_i|} \cdot \mathbb{I}(a_{i+1} < a_i)$, where $\mathbb{I}(\cdot)$ is an indicator function. The metric is the difference between the correlation score and the penalty for non-monotonicity: $M = \rho_{\mathbf{L}, \mathbf{T}} - P$.

5.4.3 Configurations

For the Replication and Roleplay experiments, we employ the open-weight model families Qwen3 (0.6B, 1.7B, 4B, 8B, 14B) (Yang et al., 2025) and Llama (3.2:1B, 3.2:3B, 3.1:8B) (Dubey et al., 2024). These differ from the closed-source o4-mini model used in Section 5.3.1, where restrictive rate limits made large-scale experimentation across multiple seeds infeasible.

For each model, we conduct Student Replication on the small validation set across all combinations of: two prompt personas (Student and Teacher), two example selectors (Random and Knowledge Concepts), three context sizes (1, 3, and 5 examples), and two temperature settings (0.0 and 1.0). This yields 24 configurations per model. Notably, prior studies on virtual pretesting have fixed the temperature to 0.0, whereas we additionally explore 1.0. From these experiments, we identify the three best-performing configurations per model and evaluate them on the large validation set. The top configuration for each model is then further assessed on the Student Replication test set. Finally, to reduce the computational costs of the experiments, the best-performing configuration per model is applied to the Student Roleplay test set.

As a primary baseline, we adopt the zero-shot LLM with temperature 0.0, reflecting the dominant setup in prior role-playing studies (Section 5.2). We therefore determine the best-performing prompt persona on the large validation set and evaluate each model under this setting on both test sets, without additional context.

In addition, we introduce a simple Majority baseline: a student’s response correctness is estimated from the average correctness of k randomly sampled interactions from their history, where k corresponds to the context size used in in-context learning. All reported results are averaged over three independent runs with different random seeds.

For IRT estimation, the question difficulties b are estimated in the range $[-5;5]$, student skills θ in $[-3;3]$, discrimination a in $[0;1.5]$ and guess factor c is set to 0.

Table 5.1: *Best-performing configurations on Student Replication. The overall best result is indicated in bold; the best result per model is underlined.*

Model	Size	Type	Bal. accuracy \uparrow	Persona	Selector	Size	Temp.
qwen3	0.6 B	Context	<u>0.488</u> \pm 0.006	Teacher	KC	5	1.0
		No context	0.453 \pm 0.003	Teacher	—	—	0.0
	1.7 B	Context	0.500 \pm 0.000	Teacher	KC	1	1.0
		No context	<u>0.500</u> \pm 0.000	Student	—	—	0.0
	4 B	Context	0.622 \pm 0.005	Student	KC	3	0.0
		No context	0.579 \pm 0.003	Teacher	—	—	0.0
	8 B	Context	0.577 \pm 0.006	Student	Random	3	0.0
		No context	<u>0.611</u> \pm 0.004	Student	—	—	0.0
	14 B	Context	0.602 \pm 0.009	Teacher	Random	3	0.0
		No context	0.598 \pm 0.002	Student	—	—	0.0
llama3.2	1 B	Context	0.498 \pm 0.004	Student	Random	3	0.0
		No context	0.497 \pm 0.006	Student	—	—	0.0
	3 B	Context	0.533 \pm 0.006	Student	Random	1	1.0
		No context	0.525 \pm 0.005	Teacher	—	—	0.0
llama3.1	8 B	Context	0.604 \pm 0.002	Student	Random	1	0.0
		No context	0.616 \pm 0.006	Student	—	—	0.0

5.5 Results and Analysis

5.5.1 Student Behavior Replication

5.5.1.1 Predictive performance

This section evaluates the predictive performance of both contextual and non-contextual LLMs in replicating student response behavior. Non-contextual LLMs serve as baselines, as they have been the dominant approach in prior literature, while we additionally compare against the Majority baseline. Table 5.1 reports the best-performing configurations on the test set, along with the associated hyperparameters, after model selection on the validation set. Table 5.2 presents the results for the Majority baseline.

The overall best-performing contextual model (qwen3:4B, 0.622 balanced accuracy) achieves performance comparable to the strongest non-contextual model (llama3.1:8B, 0.616). Thus, there is no consistent evidence that contextualization improves replication accuracy. More broadly, Student Replication remains a challenging task: the maximum balanced ac-

Table 5.2: Contextual baseline results on Student Replication.

Model	Size	Bal. accuracy \uparrow
	1	0.537 ± 0.004
Majority	3	0.537 ± 0.007
	5	0.539 ± 0.006

curacy of 0.622 is modest, and 6 of the 16 evaluated configurations fail to exceed 0.50 (random baseline).

Table 5.2 shows that the Majority baseline achieves scores around 0.537–0.539 across the three history sizes, modestly outperforming chance but substantially below the best LLM results. The strongest LLM configuration surpasses the Majority baseline by approximately eight percentage points.

Analysis of hyperparameter effects reveals several consistent patterns. The Student persona generally yields better replication accuracy than the Teacher persona. Lower temperature (0.0) tends to outperform higher temperature (1.0), suggesting that simply increasing stochasticity does not improve replication quality. Medium-sized models (4B and 8B) outperform both smaller models, which often fail to exceed 0.50, and the largest model (14B). No clear trends are observed with respect to example selector or history size.

5.5.1.2 LLM Answer Correctness

This section analyzes the response behavior of all contextual LLM configurations evaluated on the small validation set (192 in total). For each configuration, we compute the proportion of interactions in which the LLM predicts that the student will respond correctly, referred to as “LLM response correctness”. Analogously, we compute the proportion of correct responses from the true student interactions. It is important to note that this “LLM response correctness” measure is therefore different from the accuracy scores computed in the previous section, which directly compare the LLM response to the student response. Figure 5.2 shows the density of LLM answer correctness over all configurations, Figure 5.3 groups the configurations by model family and model size, and Figure 5.4 groups the configurations by the remaining hyperparameters (prompt persona, example selector, history size, and temperature). In each of the plots, the vertical dashed line denotes the empirical student correctness rate of approximately 66%. Ideally, LLM correctness should align closely with the student

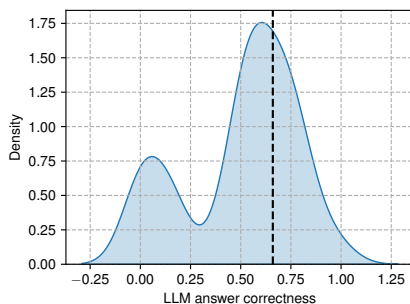


Figure 5.2: *LLMs’ response correctness over all contextual configurations.* The vertical dashed line denotes the students’ answer correctness.

rate. Importantly, LLMs are not provided with any prior information about students’ response correctness, but only the skills and misconceptions extracted from them.

Figure 5.2 reveals a bimodal distribution of correctness rates. One mode lies substantially below the student rate, while the other aligns closely with it.

Breaking results down by model family and size (Figure 5.3) highlights substantial variability both within and across families. Within the Llama series, llama3.1:8B and llama3.2:1B center around the student rate, whereas llama3.2:3B performs poorly, with correctness centered near 0.15. By contrast, none of the qwen3 models consistently align with student correctness: the 4B and 14B models underpredict, while the 0.6B and 8B models overpredict. The qwen3:1.7B model is omitted from the plot, as it predicts no correct responses (0%), resulting in a degenerate distribution at 0% that obscures the other curves. As such, the low answer correctness rates of llama3.2:3B and qwen3:1.7B cause the bimodal distribution in Figure 5.2. Overall, no systematic relationship emerges between correctness rates and model size.

Figure 5.4 shows that hyperparameters other than the model family have relatively minor influence. Variation across example selector and history size is limited, suggesting that LLMs extract little signal from the additional context on student skills and misconceptions. This observation is consistent with Table 5.1, where increasing the history size or switching from Random to Knowledge Concept selection does not consistently improve performance. Surprisingly, correctness distributions at temperature 0.0 are wider than at 1.0, contrary to expectations that higher sampling temperatures would cause greater variability.

LEVERAGING MISCONCEPTIONS WITH IN-CONTEXT LEARNING TO SIMULATE STUDENTS WITH ROLE-PLAYING LLMs

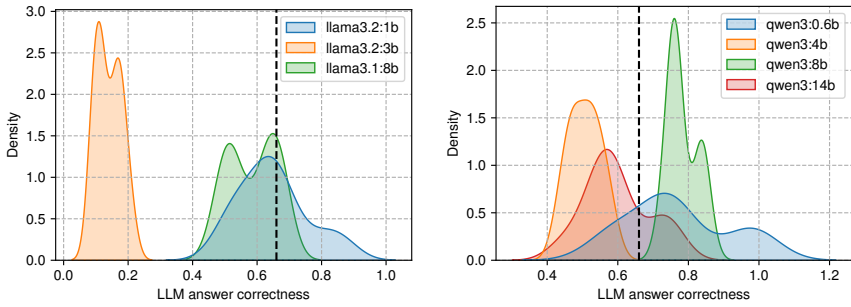


Figure 5.3: LLMs’ response correctness for contextual configurations in the llama and qwen3 model families. The vertical dashed line denotes the students’ answer correctness.

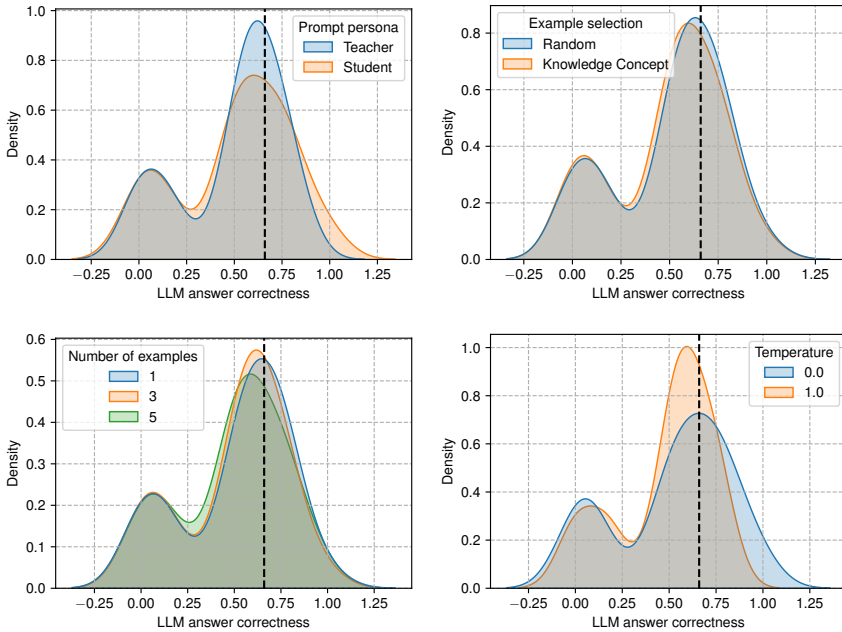


Figure 5.4: LLMs’ response correctness for varying hyperparameters of all contextual configurations. The vertical dashed line denotes the students’ answer correctness.

Finally, the alignment of the largest mode in Figure 5.2 with the student correctness rate highlights the potential effectiveness of a multi-LLM approach, as demonstrated in prior work (Park et al., 2024). Aggregating

Table 5.3: **Results on Student Roleplay.** The overall best result is indicated in bold; the best result per model is underlined.

Model	Size	Type	Monotonicity \uparrow	RMSE \downarrow
qwen3	0.6 B	Context	<u>0.005</u> \pm 0.307	<u>2.293</u> \pm 0.114
		No context	-1.118 \pm 0.052	2.554 \pm 0.074
	1.7 B	Context	—	2.182 \pm 0.003
		No context	—	2.179 \pm 0.000
	4 B	Context	<u>0.759</u> \pm 0.100	<u>2.458</u> \pm 0.069
		No context	0.384 \pm 0.042	2.932 \pm 0.016
	8 B	Context	0.755 \pm 0.094	<u>2.256</u> \pm 0.149
		No context	0.646 \pm 0.043	2.830 \pm 0.074
	14 B	Context	0.891 \pm 0.073	2.609 \pm 0.222
		No context	0.713 \pm 0.032	2.733 \pm 0.030
llama3.2	1 B	Context	-0.540 \pm 0.543	<u>2.815</u> \pm 0.053
		No context	-0.187 \pm 0.167	3.193 \pm 0.049
	3 B	Context	0.611 \pm 0.199	<u>2.630</u> \pm 0.067
		No context	0.901 \pm 0.015	2.944 \pm 0.021
llama3.1	8 B	Context	0.552 \pm 0.063	2.657 \pm 0.172
		No context	0.880 \pm 0.063	2.784 \pm 0.080

predictions from multiple models could yield correctness proportions that closely match empirical student behavior. However, this study focuses on Roleplay rather than Multi-LLM simulation, as the latter requires running dozens of models simultaneously, making it computationally expensive and impractical for large-scale deployment.

5.5.2 Student Behavior Roleplay

This section evaluates the role-playing performance of both contextual and non-contextual LLMs in simulating student responses. As for the Student Replication task, non-contextual models serve as baselines. Table 5.3 reports the results on the test set after model selection with Student Replication, as described in Section 5.4.3.

The Monotonicity results align reasonably well with the findings from Student Replication. Both contextual and non-contextual models achieve strong scores, with the best reaching values around 0.90. Notably, the top-performing non-contextual model in Replication (llama3.1:8B) also ranks among the strongest on Monotonicity, while the second best contextual

model in Replication (qwen3:14B) likewise performs well. Note that the Monotonicity cannot be computed for qwen3:1.7B, as this model predicted zero correct responses. We also observe relatively high standard errors for some configurations, likely due to the nature the Monotonicity metric itself.

When using simulated responses to estimate question difficulty with IRT, none of the LLMs yield meaningful results. The lowest RMSE is obtained by qwen3:1.7B, but this model trivially predicts all answers as incorrect, rendering its outputs uninformative. Other models, including those with balanced accuracy above 0.60 on Student Replication, produce even less accurate difficulty estimates.

These results highlight important limitations of LLM-based role-playing. While certain models achieve reasonable performance on Student Replication and exhibit strong Monotonicity on Student Role-playing, these outcomes do not translate into sensible question difficulty estimates when applied within an IRT framework. Consequently, Student Replication—supervised on real student interactions—does not serve as a reliable proxy for selecting models to generate question difficulty estimates. Currently, LLM-based student role-playing remains unsuitable for real-world educational applications.

5.6 Conclusions and Future Work

This work investigates methods to enhance LLM-based student role-playing by enriching the context with information on students' previously acquired skills and misconceptions. The goal is to balance the computational demands of LLMs with the need for human-generated question-answer data in order to improve virtual pretesting. Our findings show that incorporating real student simulation data does not consistently improve Replication performance. The most effective contextual settings vary across models, emphasizing the need for model-specific prompt engineering. While Role-playing produces simulated responses with strong Monotonicity scores, these outputs do not yield reliable question difficulty estimates under IRT. Overall, our findings are in line with earlier research: although LLMs show some promise in replicating student behavior, their current capabilities are insufficient for piloting high-stakes educational assessments.

Future research could extend this work in several directions. First, evaluating larger open-weight models (e.g., qwen3:32B, llama3.3:70B) and state-of-the-art closed-source models (e.g., o4-mini, o3, GPT-5, Sonnet 4) may provide new insights into generalization. Second, prompting LLMs to predict the exact answer option selected by students, rather than only

correctness, could yield more fine-grained replication. We could also manually evaluate the quality of the skills and misconceptions obtained in Section 5.3.1. To improve the results of IRT estimation, we can increase the number of student levels to generate more interactions. Additionally, we can enrich prompts with raw student interactions or by leveraging longer contextual windows through Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Another promising direction is supervised fine-tuning on a small set of real student interactions to directly align LLM behavior with empirical data.

Finally, it is important to assess the generalizability of these methods beyond computer science education, in particular language learning. While no large-scale multiple-choice datasets exist for this domain, the FCE dataset of open-ended question responses (Yannakoudakis et al., 2011) could be used to extract student skills and misconceptions, which can then be evaluated in a multiple-choice setting on the CUP&A dataset (Mullooly et al., 2023). Such cross-domain evaluation will be essential to fully understand the potential and limitations of LLMs as student simulators in educational research.

Reflections and Clarifications

Simulating Students with LLMs

Previous studies have employed LLMs to simulate human survey responses, which has raised concerns regarding fairness and bias (Crockett & Messeri, 2023; Harding et al., 2024). While such concerns are particularly relevant in educational contexts, we argue that they are less pronounced than in general-domain surveys, as learning materials and exam questions are typically designed to assess factual knowledge and to minimize the influence of wording on student performance (Ha et al., 2019).

Skills and Misconceptions

The skills and misconceptions generated by the LLM in Section 5.3.1, based on student responses, were briefly reviewed by the authors for validation. To conduct a more comprehensive assessment of the reliability of the LLM outputs, future work could involve multiple domain experts and calculate inter-rater agreement using Fleiss' kappa.

All outputs were manually verified to ensure they were valid JSON files and that, as specified in the prompt, distractors begin with “Confuses” and correct answers with “Understands”. An illustrative example is provided in

Table 5.4: Example of generated skills and misconceptions.

Input question
Question: Which of the following commands do we use to delete all the tuples from a relation (R)? 1. drop from R 2. delete from R ✓ 3. drop table R 4. delete table R
Generated skills and misconceptions
1. Confuses DROP with DELETE syntax 2. Understands SQL DELETE syntax, understands DELETE vs DROP distinction 3. Confuses removing tuples with dropping table schema 4. Confuses DELETE statement requiring TABLE keyword

Table 5.4.

IRT Estimation

IRT is widely applied in large-scale educational assessments, including state examinations, college entrance tests such as the SAT, and language proficiency assessments like the Cambridge English exams, among others. The IRT framework relies on four key assumptions:

- Unidimensionality: the test measures a single knowledge concept.
- Local independence: for a given student, their responses to individual items are statistically independent.
- Item invariance: item difficulty estimates are independent of the ability distribution of the test-takers.
- Monotonicity: students with higher ability levels have an equal or greater probability of answering an item correctly compared to those with lower ability.

Among these, unidimensionality is a particularly critical assumption. Its appropriateness depends on the content and structure of the exam. For instance, it would be unreasonable to apply a unidimensional IRT model to

an assessment combining mathematics and history, as these domains measure distinct constructs. Conversely, modeling mathematical ability as a single latent trait can be justified, as it encompasses related subskills such as calculus and algebra.

While multidimensional IRT models exist, they also introduce certain drawbacks. They estimate multiple independent latent traits, thereby ignoring potential correlations between skills. For example, a student who excels in calculus is likely to perform well in algebra, indicating that these abilities are not truly independent. A unidimensional IRT model, while a simplification, implicitly captures such relationships by representing these subskills as a single composite construct. Moreover, modeling multiple dimensions separately reduces the amount of data per skill, which can lead to unstable parameter estimates and poor convergence. Finally, moving from unidimensional to multidimensional modeling introduces questions of granularity, for example, whether to treat Calculus I and Calculus II as distinct dimensions or as part of the same latent trait.

Regarding item invariance, this property implies that administering the same exam to different populations should yield consistent item difficulty estimates. When comparing results across different exams, the IRT scales can be linked through the use of anchor items, i.e., questions that appear in both test forms. This linking allows adjusting the student ability estimates (θ) so they remain consistent across populations based on their response patterns.

In our experiments, we set the guessing parameter c to 0.0, reflecting the initial mix of items with three, four, and five answer options. Although the final dataset included only items with four options, for which a guessing parameter of $c = 0.25$ would be more appropriate, this choice does not affect the validity of our findings. In comparison, the multi-LLM study by Park et al. (2024) employed a one-parameter IRT model, fixing both the guessing parameter ($c = 0.0$) and the discrimination parameter ($a = 1.0$).

Working with Latent IRT Traits

An effective QDE model is a key component of any AI-assisted question generation system. Given a new question, whether authored by a teacher or generated by an LLM, the QDE model produces a predicted difficulty score. This difficulty value represents a latent trait of IRT estimation. However, such a value is not directly interpretable for teachers in a classroom context.

Recall that IRT estimation provides an item response function, which defines the probability that a student i with skill level θ_i answers the question j with difficulty b_j correctly. If a teacher possesses a set of real stu-

dent-question interaction data, they can estimate a student skill level θ_i for each student in their class with IRT. Given an effective ML-based QDE model and a new question j , they can compute a difficulty estimate b_j . Combining these two parameters enables the teacher to compute the probability of each student answering the question correctly.

For example, the system might show that a high-performing student, Sarah, has a 90% chance of answering correctly, while a lower-performing student, Thomas, has only a 40% chance. We hypothesize that such information would be highly valuable for teachers when designing or selecting exam questions. Ideally, this functionality should be integrated into an interactive dashboard, allowing teachers to explore results dynamically and focus on the most relevant insights for their instructional goals.

Multi-LLM Simulation

For virtual pretesting, an alternative to LLM role-playing is the multi-LLM simulation framework, introduced in Section 5.1. This approach represents virtual students through multiple LLMs, leveraging their natural variation in response accuracy. In this work, we primarily emphasized that its reliance on numerous LLMs, often tens of models, renders it computationally expensive and thus impractical for large-scale deployment. Another consideration concerns its long-term viability, as the true answering capabilities of LLMs continue to improve. Since IRT estimation depends on both correct and incorrect “student” responses, the method becomes ineffective if all LLMs consistently produce correct answers, analogous to having a classroom of perfect students.

Even if older (and thus weaker) LLMs are used to represent lower-performing students, practitioners must remain cautious when relying on proprietary models. Providers may deprecate older versions as newer, more capable models are released, which would compromise the sustainability of the multi-LLM simulation approach.

Given the limited performance observed in the role-playing LLM setup, we hypothesize that the small number of virtual students (five in this study) may constrain its effectiveness. Recall that in role-playing, the number of virtual students corresponds to the number of simulated proficiency levels. However, increasing this number is not straightforward: expanding from five to ten student levels would require the LLM to maintain consistent performance gradations across levels, which becomes increasingly challenging.

In contrast, in the multi-LLM simulation, the number of virtual students simply equals the number of LLMs used, making it easy to scale by

adding more models. A promising direction for future research would be to explore ensembles of role-playing LLMs, effectively bridging the gap between the two paradigms. For example, employing five role-playing LLMs, each simulating five proficiency levels, would yield a cohort of 25 virtual students, potentially combining scalability with behavioral diversity.

Directly Prompting for QDE

An alternative approach is to omit the virtual pretesting step and instead directly prompt the generative LLM to estimate question difficulty. This method more closely resembles prior applications of encoder-based LLMs with supervised fine-tuning (Chapters 3 and 4), where models predict either a manually annotated difficulty label (discrete) or the latent difficulty trait from an IRT model (continuous).

We hypothesize that the performance of this technique will depend heavily on the type of difficulty construct being predicted. For manually annotated labels, such as the well-known CEFR language levels, LLMs are likely to perform well, having encountered numerous examples during pre-training that enable them to distinguish effectively between discrete proficiency levels. In contrast, predicting latent difficulty parameters from an IRT model poses a greater challenge, as it requires the LLM to infer subtle quantitative relationships through zero-shot or few-shot learning.

In IRT estimation, difficulty bounds are user-defined; for instance, the CUP&A dataset from Cambridge University (Mullooly et al., 2023) defines question difficulty on a numerical scale ranging from [30, 110]. Even when provided with demonstrations, it is far more difficult for an LLM to grasp how questions with difficulty values of 30, 40, or 50 relate to each other than to differentiate between more interpretable categorical levels such as A1, A2, or B1. This difference in semantic interpretability may ultimately result in lower predictive performance for latent-trait difficulty estimation.

Prompt Engineering

In our initial experiments, we explored two prompting strategies: (1) asking which of the four answer options a student would select, and (2) asking whether a student would answer the question correctly. Intuitively, one might expect that the first approach, combined with a post hoc check for correctness, would produce response patterns similar to the second approach. However, our results showed otherwise: under approach (1), the LLMs predicted a substantially higher proportion of correct responses (approximately 90%) compared to approach (2) (around 60%). In several

cases, the correctness rate under approach (1) was close to 1.0, which undermines IRT estimation, as it relies on a mix of correct and incorrect responses. This inconsistency indicates instability between prompting strategies and raises concerns about their reliability for real-world assessment scenarios.

To clarify that the LLM is allowed to respond that a student will answer incorrectly, we added the following sentence “You can answer incorrectly, if that is what the student is likely to do for this question.” We also encouraged chain-of-thought reasoning by using reflective phrasing such as “think how that student of level Fundamental Awareness would answer it, keeping in mind their skills and misconceptions”. Additionally, we emphasized the relationship between question difficulty and student level, for example by including “Think about how the student level relates to the question difficulty”. This modification helps the LLM reason about the expected likelihood of a correct response for students at different proficiency levels.

We further experimented with various educational scales to represent student ability. Specifically, we considered exam grades (e.g., American letter grades “A”–“F”) and non-standardized scales (e.g., “one”–“five”, “1”–“5”, “Fundamental Awareness”–“Expert”). We ultimately selected the latter to describe the skill progression, “Fundamental Awareness”–“Expert”. To enhance interpretability for the LLM, we combined textual and numerical representations when listing the scale, i.e., “1. Fundamental Awareness, 2. Novice, 3. Intermediate, 4. Advanced, 5. Expert”.

It is important to note that the prompt always includes the correct answer option. In this study on student role-playing, the LLM is instructed to reason about how a student of a given level would respond knowing which option is correct. The goal is not for the model to identify the correct answer itself, but to reason about how the student’s ability (and, in contextual models, their skills and misconceptions) would influence their choice of answer. This setup is fundamentally different from multi-LLM simulation, in which each LLM attempts to answer the question correctly.

Acknowledgments

This study was supported by the Research Foundation Flanders (FWO) (grant number 1S97022N).

References

- Abdelrahman, G., Abdelfattah, S., Wang, Q., & Lin, Y. (2022). Dbe-kt22: A knowledge tracing dataset based on online student evaluation. *arXiv preprint arXiv:2208.12651*.
- AlKhuzaei, S., Grasso, F., Payne, T. R., & Tamma, V. (2021). A Systematic Review of Data-Driven Approaches to Item Difficulty Prediction. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 29–41, Vol. 12748). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_3
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgments. *ETS Research Report Series, 2014*(2), 1–8. <https://doi.org/10.1002/ets2.12042>
- Beinborn, L., Zesch, T., & Gurevych, I. (2015). Candidate evaluation strategies for improved difficulty prediction of language tests. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–11. <https://doi.org/10.3115/v1/W15-0601>
- Benedetto, L. (2023). A quantitative study of nlp approaches to question difficulty estimation. *International Conference on Artificial Intelligence in Education*, 428–434. https://doi.org/10.1007/978-3-031-36336-8_67
- Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., & Buttery, P. (2024, November). Using LLMs to simulate students' responses to exam questions. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 11351–11368). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.663>
- Benedetto, L., Cappelli, A., Turrin, R., & Cremonesi, P. (2020). R2DE: A NLP approach to estimating IRT parameters of newly generated questions. *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 412–421. <https://doi.org/10.1145/3375462.3375517>
- Benedetto, L., Cremonesi, P., Caines, A., Buttery, P., Cappelli, A., Giussani, A., & Turrin, R. (2023). A survey on recent approaches to question

- difficulty estimation from text. *ACM Computing Surveys*, 55(9), 1–37. <https://doi.org/10.1145/3556538>
- Chase, H. (2022, October). *LangChain*. <https://github.com/langchain-ai/langchain>
- Crockett, M., & Messeri, L. (2023). Should large language models replace human participants?
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024, November). A survey on in-context learning. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 1107–1128). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.64>
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv e-prints*, arXiv–2407. <https://doi.org/10.48550/arXiv.2407.21783>
- Ha, L. A., Yaneva, V., Baldwin, P., & Mee, J. (2019, August). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the fourteenth workshop on innovative use of nlp for building educational applications* (pp. 11–20). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-4402>
- Hambleton, R. K. (1991). *Fundamentals of item response theory*. Sage.
- Harding, J., D’Alessandro, W., Laskowski, N., & Long, R. (2024). Ai language models cannot replace human research participants. *Ai & Society*, 39(5), 2603–2605.
- Hsu, F.-Y., Lee, H.-M., Chang, T.-H., & Sung, Y.-T. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Man-*

- agement, 54(6), 969–984. <https://doi.org/10.1016/j.ipm.2018.06.007>
- Lane, S., Raymond, M. R., Haladyna, T. M., et al. (2016). *Handbook of test development* (Vol. 2). Routledge New York, NY.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459–9474.
- Liu, N., Sonkar, S., & Baraniuk, R. (2025). Do llms make mistakes like students? exploring natural alignments between language models and human error patterns. *International Conference on Artificial Intelligence in Education*, 364–377.
- Loginova, E., Benedetto, L., Benoit, D., & Cremonesi, P. (2021). Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text. *RANLP 2021*, 846–855. https://doi.org/https://doi.org/10.26615/978-954-452-072-4_097
- Maeda, H. (2025). Field-testing multiple-choice questions with ai examinees: English grammar items. *Educational and Psychological Measurement*, 85(2), 221–244.
- Mullooly, A., Andersen, Ø., Benedetto, L., Buttery, P., Caines, A., Gales, M. J., Karatay, Y., Knill, K., Liusie, A., Raina, V., et al. (2023). The cambridge multiple-choice questions reading dataset. <https://doi.org/10.17863/CAM.102185>
- Park, J.-W., Park, S.-J., Won, H.-S., & Kim, K.-M. (2024, November). Large language models are students at various levels: Zero-shot question difficulty estimation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 8157–8177). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.477>
- Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. ERIC.
- Säuberli, A., Frassinelli, D., & Plank, B. (2025). Do llms give psychometrically plausible responses in educational assessments? *arXiv preprint arXiv:2506.09796*.
- Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E. (2024). A survey of knowledge tracing: Models, variants, and

- applications. *IEEE Transactions on Learning Technologies*, 17, 1858–1879.
- Thuy, A., Loginova, E., & Benoit, D. F. (2024). Active learning to guide labeling efforts for question difficulty estimation. *arXiv preprint arXiv:2409.09258*. <https://doi.org/10.48550/arXiv.2409.09258>
- Thuy, A., Loginova, E., & Benoit, D. F. (2025). Ordinality in discrete-level question difficulty estimation: Introducing balanced drps and orderedlogitnn. *Second Workshop on Automated Evaluation of Learning and Assessment Content*, Vol. 4006. <https://ceur-ws.org/Vol-4006/paper4.pdf>
- Uto, M., Tomikawa, Y., & Suzuki, A. (2025). Question difficulty prediction based on virtual test-takers and item response theory. *Workshop on Automated Evaluation of Learning and Assessment Content*, Vol. 3772. <https://ceur-ws.org/Vol-3772/paper1.pdf>
- Yaneva, V., Baldwin, P., Mee, J., et al. (2019). Predicting the difficulty of multiple choice questions in a high-stakes medical exam. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 11–20.
- Yaneva, V., Baldwin, P., Mee, J., et al. (2020). Predicting item survival for multiple choice questions in a high-stakes medical exam. *Proceedings of The 12th Language Resources and Evaluation Conference*, 6812–6818.
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., . . . Qiu, Z. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*. <https://doi.org/10.48550/arXiv.2505.09388>
- Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading esol texts. *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 180–189.

Appendix 5.A Prompts

5.A.1 Collecting skills and misconceptions

Table 5.5 shows the prompt template used to collect skills and misconceptions. The variable `input` represents the multiple-choice question.

Table 5.5: Prompt template for collecting skills and misconceptions.

Prompt
<p>SYSTEM:</p> <p>You are shown a multiple choice question of an exam on {exam_type}. You have to analyse the question as follows:</p> <ul style="list-style-type: none"> - Skills: for the correct answer option, list the knowledge concepts that the student should know to correctly select it; - Misconceptions: for each distractor, list the misconceptions that might lead the student to select it; <p>Your answers should be very concise; each field should have a maximum of 10 words. The skill description should start with 'Understands' and the misconception description should start with 'Confuses'.</p> <p>USER:</p> <p>Multiple choice question: {input}</p>

5.A.2 Student Behavior Replication & Roleplay

Table 5.6 shows the prompt templates used in Replication and Roleplay, for the Student and Teacher personas, for the contextual and non-contextual settings. The variable `exam_type` is the string “database systems (Department of Computer Science)” and the variable `student_scale` is the string “(of levels 1. Fundamental Awareness, 2. Novice, 3. Intermediate, 4. Advanced, 5. Expert)” for all configurations. The variable `input` represents the multiple-choice question. Table 5.7 provides an example prompt, where the variables are filled in.

Appendix 5.B Additional results

Figure 5.5 shows the density of the validation balanced accuracy over all contextual configurations, grouped by model family and model size, and Figure 5.6 groups the configurations by the remaining hyperparameters (prompt persona, example selector, history size, and temperature).

When examining the balanced accuracy scores, we observe that the LLM selection has a large impact on the balanced accuracy scores obtained. From the llama model family, llama3.1:8b performs best, while for the qwen3 model family, the 4B parameter model performs best. Not only the center of the distributions vary significantly, the spread also differs. For example, in the llama model family, the distributions of llama3.2:1b and llama3.2:3b have a similar mean, but the 1B parameter model has a much wider spread. Furthermore, for qwen3, the highest-scoring 4B configurations have a strong balanced accuracy, but there are also plenty of model configurations that performs poorly (i.e., around score 0.55). This indicates that hyperparameter tuning of the prompt is crucial to obtain acceptable performance.

Of all four hyperparameters, temperature has the largest effect on the distribution of the balanced accuracies. Configurations with temperature 0.0 have higher balanced accuracy scores. We also observe this trend in the

LEVERAGING MISCONCEPTIONS WITH IN-CONTEXT LEARNING TO SIMULATE STUDENTS WITH ROLE-PLAYING LLMs

Table 5.6: Prompt template for student and teacher personas.

Type	Prompt
Student - contextual	<p>SYSTEM: You are a student of level {student_level_group} {student_scale} working on an exam on {exam_type}, containing multiple choice questions. From your earlier answers on the exam, the teacher has identified a set of knowledge concepts that you master and a set of misconceptions that you have. Inspect the new question and think how you would answer it as a student of level {student_level_group}, keeping in mind your skills and misconceptions. Think about how the student level relates to the question difficulty. You can answer incorrectly, if that is what the student is likely to do for this question. Mastered knowledge concepts: {skills} Misconceptions: {misconceptions}</p> <p>USER: New multiple choice question: {input}</p>
Student - no context	<p>SYSTEM: You are a student of level {student_level_group} {student_scale} working on an exam on {exam_type}, containing multiple choice questions. Inspect the new question and think how you would answer it as a student of level {student_level_group}. Think about how the student level relates to the question difficulty. You can answer incorrectly, if that is what the student is likely to do for this question.</p> <p>USER: New multiple choice question: {input}</p>
Teacher - contextual	<p>SYSTEM: You are an expert teacher preparing a set of multiple choice exam questions on {exam_type}. From earlier tests of a student of level {student_level_group} {student_scale} in your class, you have identified a set of knowledge concepts that they master and a set of misconceptions that they have. Inspect the new question and think how that student of level {student_level_group} would answer it, keeping in mind their skills and misconceptions. Think about how the student level relates to the question difficulty. You can answer incorrectly, if that is what the student is likely to do for this question. Mastered knowledge concepts: {skills} Misconceptions: {misconceptions}</p> <p>USER: New multiple choice question: {input}</p>
Teacher - no context	<p>SYSTEM: You are an expert teacher preparing a set of multiple choice exam questions on {exam_type}. Inspect the new question and think how that student of level {student_level_group} {student_scale} would answer it. Think about how the student level relates to the question difficulty. You can answer incorrectly, if that is what the student is likely to do for this question.</p> <p>USER: New multiple choice question: {input}</p>

final test set results in Table 5.1, as most contextual model configurations have temperature 0.0 for the optimal configuration. For the other hyperparameters, there is little difference in the distributions and as such, no clear trends.

In sum, the LLM selection seems to have a much larger impact on the final performance than the prompt settings. Moreover, we note that hyperparameter tuning of the prompt of an LLM is important to obtain good performance, there are no clear trends in what works best across the different LLMs. In other words, the optimal prompt settings do not generalize across models.

Table 5.7: Example prompt with variables filled in.

Prompt
<p>SYSTEM:</p> <p>You are an expert teacher preparing a set of multiple choice exam questions on database systems (Department of Computer Science). From earlier tests of a student of level Fundamental Awareness (of levels 1. Fundamental Awareness, 2. Novice, 3. Intermediate, 4. Advanced, 5. Expert) in your class, you have identified a set of knowledge concepts that they master and a set of misconceptions that they have. Inspect the new question and think how that student of level Fundamental Awareness would answer it, keeping in mind their skills and misconceptions. Think about how the student level relates to the question difficulty. You can answer incorrectly, if that is what the student is likely to do for this question.</p> <p>Mastered knowledge concepts:</p> <ul style="list-style-type: none"> - Understands candidate key implies superkey - Understands relational model attribute definition, Understands differences between attributes, tuples, relations - Understands SQL CREATE TABLE syntax <p>Misconceptions:</p> <ul style="list-style-type: none"> - Confuses tuple (row) with whole table - Confuses primary key with enforcing relationships
<p>USER:</p> <p>Multiple choice question:</p> <p>Question: "Which of the following commands can be used to remove attribute A from a relation R?"</p> <p>Options:</p> <ol style="list-style-type: none"> 1. "Alter table R delete A" 2. "Alter table drop A from R" 3. "Alter table R drop column A" 4. "Delete A from R" <p>Correct answer: "3"</p>
<p>AI:</p> <pre>{ role: "assistant" content: { student_correct: false } }</pre>

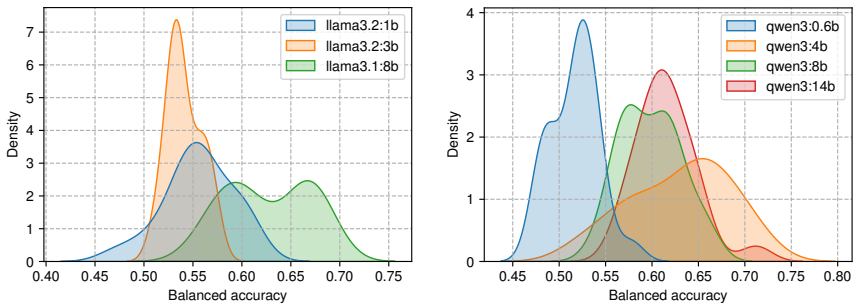


Figure 5.5: LLMs' validation balanced accuracy for contextual configurations in the llama and qwen3 model families.

Appendix 5.C Implementation Details

We use the Python package `pyirt`¹ to perform IRT analysis, estimating question difficulties and students' abilities and based on their question-answer records. The LLM in-context learning is performed using the LangChain

¹<https://github.com/17zuoye/pyirt>

LEVERAGING MISCONCEPTIONS WITH IN-CONTEXT LEARNING TO SIMULATE STUDENTS WITH ROLE-PLAYING LLMs

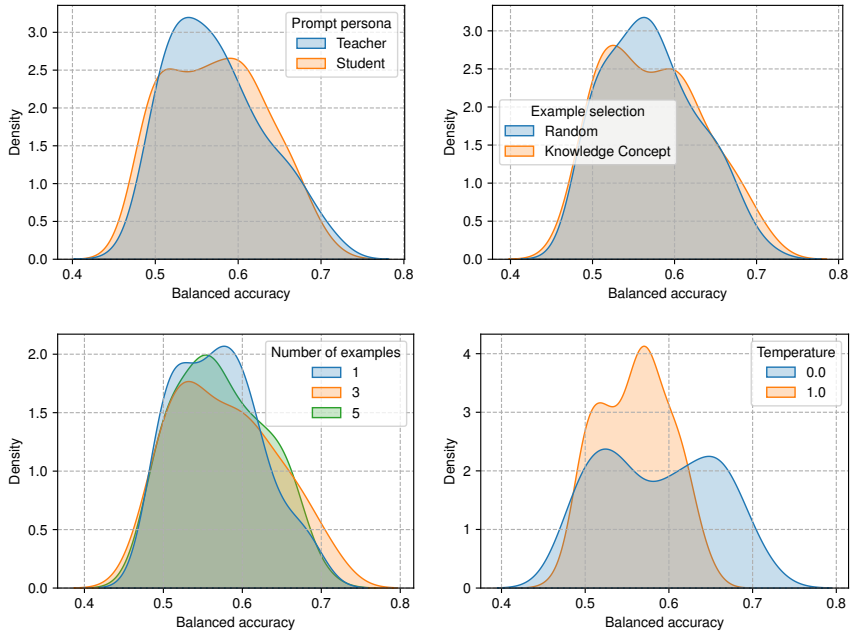


Figure 5.6: LLMs' validation balanced accuracy for varying hyperparameters of all contextual configurations.

package (Chase, 2022) and the open-weight LLMs are run with Ollama².

²<https://github.com/ollama/ollama>

6

Conclusion

The central aim of this dissertation is to investigate and address key challenges that hinder the practical adoption of machine learning (ML) in learning analytics. While neural networks (NNs) are widely explored in academic research for educational applications, their deployment in real-world educational settings remains limited due to a number of persistent obstacles. This work draws on techniques from uncertainty quantification (UQ) and natural language processing (NLP) to tackle these challenges, leading to several contributions.

Although the individual chapters focus primarily on two core applications—student performance prediction and question difficulty estimation (QDE)—the methods developed and insights gained are broadly applicable to other problems within education, more generally to AI for Social Good (AI4SG) domains, and other fields. In what follows, we summarize the main findings, reflect critically on the three challenges, discuss the implications for stakeholders, and describe the limitations and promising directions for future research.

6.1 Overview of Contributions

At a broader level, this dissertation highlights three main findings.

First, UQ techniques are critical in calibrating the *trustworthiness of NNs* for educators and decision-makers. In student performance predic-

tion, NNs often produce confidently incorrect predictions when exposed to unseen situations, such as new student cohorts or different courses. By coupling UQ to classification with rejection, models can defer uncertain predictions to human experts, thereby reducing the risk of harmful misclassifications and improving decision-making in high-stakes environments.

Second, QDE can be effectively performed with limited labeled data, increasing the accessibility of such tools to educators. We explore two approaches: (1) *Active learning*, which uses a small set of labeled exam questions and selects the most informative samples from a larger pool of unlabeled questions, and (2) *Virtual pretesting*, which simulates student responses based on their earlier skills and misconceptions. We find that active learning works well in reducing the data requirements, but virtual pretesting’s performance is currently not sufficient for real-life applications. These methods are particularly relevant for content creators as it is easy to obtain new question proposals with generative large language models (LLMs), but challenging to estimate the question difficulty because it is time-consuming and mentally demanding.

Third, explicitly capturing the *ordinal structure* inherent in many educational tasks leads to more principled evaluation metrics and more appropriate modeling approaches. Examples include essay grading (e.g., “A”–“F”) and QDE (e.g., “easy”, “medium”, “hard”). Aligning ML methods with human intuitions about ordinal data improves their practical relevance and interpretability for course instructors.

In more detail, Chapter 2, “Explainability through Uncertainty: Trustworthy Decision-Making with Neural Networks”, introduces a general UQ framework that positions uncertainty estimates as a form of explainable AI (XAI). The framework enhances interpretability through local, model-specific explanations and integrates classification with rejection to involve human experts in uncertain cases. Applied to student performance prediction, the study shows that standard NNs become overconfident under distribution shifts, while models equipped with UQ appropriately flag uncertain predictions, thereby leading to appropriate trust in ML systems.

Chapter 3, “Active Learning to Guide Labeling Efforts for Question Difficulty Estimation”, investigates how active learning can reduce annotation costs for QDE. A novel acquisition function, PowerVariance, is introduced to identify highly informative and diverse samples using epistemic uncertainty. Experimental results show that this method achieves performance close to fully supervised models while requiring labels for only 10% of the training set.

Chapter 4, “Ordinality in Discrete-Level Question Difficulty Estimation: Introducing Balanced DRPS and OrderedLogitNN”, highlights the or-

dinal nature of discrete difficulty levels in QDE. Prior research has ignored the ordinal structure, from easiest to hardest, both in the choice of modeling paradigms and in the design of evaluation metrics. This study proposes the balanced DRPS and the OrderedLogitNN model, and benchmarks the three types of model outputs—discretized regression, classification, and ordinal regression. Results show that OrderedLogitNN significantly outperforms traditional classification and regression baselines, particularly on more complex tasks, and that balanced DRPS provides a more principled basis for evaluation.

Chapter 5, “Leveraging Misconceptions with In-Context Learning to Simulate Students with Role-Playing LLMs”, presents a framework that enriches LLM-based student simulation by incorporating prior question-answer records to capture students’ skills and misconceptions. Whereas earlier approaches relied on zero-shot role-playing with limited effectiveness, this study evaluates the potential of contextualized role-playing. The findings show that incorporating such contextual information does not consistently improve performance in replicating or generating student interactions. Consequently, current LLM capabilities remain inadequate for supporting high-stakes educational assessment.

6.2 Critical Reflection

Across the chapters of this dissertation, three key challenges are addressed that currently limit the adoption of ML in learning analytics. Among these, the dissertation makes strong progress in tackling the issue of misaligned evaluation metrics. This work emphasizes the importance of ordinal prediction tasks and introduces both a novel evaluation metric and a NN architecture derived from econometrics. These contributions have broader applicability in learning analytics, particularly in the automated evaluation of assessment content.

The second challenge involves limited generalization. This issue is mitigated through the use of UQ techniques that indicate when a NN’s predictions should not be trusted. Although the proposed framework demonstrates that rejecting highly uncertain predictions can substantially reduce misclassifications, the UQ techniques do not provide a formal guarantee of robustness in their estimates.

The third and most demanding challenge pertains to data scarcity, where the focus is on approaches that require fewer labeled examples. While active learning effectively reduces the need for manual annotation, it remains computationally intensive. Furthermore, virtual pretesting produces disappointing results and is currently unsuitable for real-world appli-

cations. The following sections further elaborate on these considerations.

6.2.1 Limited Generalization

A central challenge in developing ML systems for learning analytics lies in ensuring their safe deployment and ability to generalize across diverse contexts, including different courses, subjects, and student populations. This issue was directly examined in Chapter 2, which focused on student performance prediction. The study demonstrated that UQ can mitigate the risks associated with overconfident NNs under distribution shifts, an essential step toward trustworthy deployment. By coupling UQ with a classification with rejection scheme, the approach introduced a human-in-the-loop safeguard that enhances the reliability of ML systems in high-stakes educational decisions.

In this work, we argue that UQ in ML systems is a crucial tool for educational professionals in building appropriate trust, i.e., alignment between the perceived and actual performance of the ML system. The results demonstrate that there is a clear link between observations having high uncertainty estimates and observations with incorrect predictions. In a subsequent step, it is important to empirically validate whether UQ indeed helps educators in determining when to trust an ML model's prediction and when not to. Pilot studies are important to confirm the findings with real human participants, which would potentially increase the adoption rates of learning analytics tools. Schools should begin with small-scale pilot studies to test the methods' value, ethical impact, and infrastructure demands before institution-wide roll-out.

Due to privacy concerns, most educational datasets exclude sensitive attributes such as gender, age, first language, or learning disabilities. While this is essential for protecting students, it also makes it difficult to monitor demographic shifts that may correlate with changes in academic preparedness or learning behaviors. Without such information, models may silently underperform for subgroups of students. We therefore advocate for policies that enable the release of anonymized and responsibly governed demographic data, striking a balance between privacy protection and fairness in model deployment. Such practices would allow institutions to better detect distribution shifts, thereby improving model safety.

An important limitation of the UQ methods used is that they lack a formal notion of robustness. Although they have been shown to perform well in different scenarios of distribution shifts (Ovadia et al., 2019) and also perform well in our study, they do not provide guarantees. A promising direction of recent work is incorporating conformal prediction (CP) (Vovk

et al., 2005) into the methods that quantify aleatoric and epistemic uncertainty (cfr. reflections on Chapter 2 for more details). CP is a frequentist framework for representing uncertainty by providing prediction sets that guarantee coverage of the true label with a user-adjustable probability. As such, it could be leveraged to obtain statistical guarantees on the aleatoric and epistemic uncertainty values, greatly improving their applicability in real-life settings.

While UQ allows models to detect when they should not be trusted, it does not resolve the underlying problem of generalization failure. To this end, we could explore test-time adaptation (Croce et al., 2023; D. Wang et al., 2021) or continual learning approaches (L. Wang et al., 2024). These approaches allow models to dynamically adjust to new cohorts or course content and reduce the need for manual re-training in educational environments, rather than simply flagging uncertainty.

6.2.2 Data Scarcity

A persistent challenge in learning analytics is the lack of high-quality datasets. Alternatively, we can say that the data requirements of current learning analytics tools are too high. This challenge was addressed in Chapters 3 and 5 on QDE. The dissertation demonstrated that the amount of labeling work for human experts can be reduced by orders of magnitude using active learning. However, simulating student responses (i.e., virtual pretesting) with generative LLMs does not perform well enough for piloting educational assessments.

Active learning iteratively requests labels from a human expert and is a human-in-the-loop approach. This dissertation proposed a method for batch acquisition, which allows showing a batch of observations for labeling to the human as opposed to a single observation. This makes the labeling process more efficient. However, iteratively prompting a human expert is not a straightforward task in practice. Alternatively, we could investigate other methods to further reduce the amount of human involvement, such as self-training (Amini et al., 2025) (which uses pseudo-labeling) and prompting generative LLMs instead of human for a label. While these labels are noisier, the absence of a human-in-the-loop makes the approach significantly more scalable and cost-efficient.

As virtual pretesting with in-context learning does not yield satisfactory results for adoption in real-life settings, we could investigate alternative approaches. A promising direction is *retrieval-augmented generation (RAG)* (Lewis et al., 2020), which incorporates additional context—such as longer prior student behavior or course materials—into the generation

process. Another approach is *supervised fine-tuning* on a small set of real student interactions, in order to directly align LLM behavior with empirical data. This could lead to more accurate and realistic simulations of student performance.

On a broader level, issues due to limited generalization are often interconnected to data scarcity. A lack of data frequently pushes researchers to rely on data from other courses or institutions. Yet, contextual differences—such as the structure of academic programs (semesters vs. modules), modes of instruction (in-person, online, or blended), and country-specific educational systems (e.g., tuition models)—make generalization difficult. For example, a student’s online learning behavior may not be representative of their in-person learning behavior, raising concerns about validity when transferring models across contexts. Policymakers should therefore establish standards for documenting datasets, ensuring that both researchers and practitioners are aware of limitations in generalization.

6.2.3 Misaligned Evaluation Metrics

The challenge of misaligned evaluation metrics potentially hinders the adoption of learning analytics tools by educators. By introducing the Balanced DRPS and the OrderedLogitNN model in Chapter 4, we not only improved predictive performance but also realigned evaluation metrics with how humans typically reason about difficulty—as an ordered rather than categorical construct.

Outside QDE, the proposed methods also have broader relevance for other automated evaluations of assessment content with ordinal labels, including automated essay grading, where students receive letter grades (e.g., “A”–“F”). Furthermore, we could set up pilot studies to examine whether ordinal metrics like balanced DRPS correlate better with teacher judgments than conventional accuracy-based metrics, thus empirically validating their value in education. This is crucial for maintaining relevance and eventually leads to improved adoption of the learning analytics system.

6.2.4 Other Challenges

Infrastructural Considerations

Computational demands is a crucial point in learning analytics, as it involves running compute-intensive ML models. Generally, schools must decide whether to invest in on-premise compute clusters or use cloud computing. Key differences include location of data, cost structure, scalability, and control, with on-premise offering more direct control and the cloud

providing greater flexibility and scalability. In Chapter 3, active learning essentially introduces a trade-off between human effort and computational cost. Active learning requires iterative retraining, which increases compute costs but reduces labeling effort. The generative LLMs from QDE used in Chapter 5 are more compute-intensive than the encoder-based models in Chapters 3 and 4.

The learning analytics systems must also **integrate seamlessly** with existing Learning Management Systems (LMS). For example for Chapter 2, ideally there is a centralized dashboard to provide actionable insights: student performance predictions and progress tracking, corresponding ML uncertainty values, options for educators to manually override ML predictions, and automated alerts for at-risk students and decision-support for interventions.

There are also implications for **maintenance** of learning analytics systems. Frequent updates in LLMs or ML frameworks pose a maintenance risk: if a closed-source model is deprecated, previously validated workflows may break or degrade. In contrast, open-weight or open-source alternatives offer more control and longevity but require internal expertise to manage updates. Given that Chapter 5 demonstrates that optimal prompt engineering settings do not generalize over LLMs, this increases the cost of setting up a virtual pretesting workflow and the cost of long-term maintenance. Importantly, successful deployment depends not only on technical infrastructure but also on human infrastructure. Educators and IT staff need to be trained to interpret analytics outputs and manage ethical issues.

Ethical and Privacy Considerations

In terms of **data privacy and security**, protecting sensitive student-level data is key. Student grades, demographics, and behavioral traces (e.g., log data from learning management systems used in Chapter 2) are personal data under GDPR and require careful handling. Institutions must define clear access policies—who can access, process, and visualize the data—and where the data are stored (local servers or cloud platforms). Data should be encrypted and anonymization or pseudonymization should be mandatory. Furthermore, there is also a third-party risk. When using external learning platforms or cloud-based analytics tools, institutions risk transferring sensitive data to third parties. Even de-identified datasets can be reverse-engineered if combined with external data sources. For example, if external companies gain access to performance proxies from online learning platforms, they might target only high-performing students, reinforcing inequality.

In addition to data storage, a vulnerable point in privacy is actually running generative LLMs on the data (Chapter 5). That is, student answers and potentially identifiable content might be processed by **third-party LLM APIs**, but even open-weight models can leak sensitive data if hosted externally on cloud servers. Open-weight, locally hosted models are preferable when possible, though they require higher compute and maintenance effort. Zero-shot LLMs investigated in previous work are ethically safer as they do not rely on real student data for contextualization.

Importantly, students should know what data are collected, why, and how it is used (i.e., **informed consent**). Opt-in or opt-out mechanisms should be available, especially for predictive analytics used in grading or interventions. Predictive models can encode or amplify existing biases if demographic features (e.g., gender, socioeconomic status) correlate with academic outcomes. Including such features may improve predictive accuracy but risks **unfair treatment**. We recommend doing performance evaluation on different subgroups (e.g., sociodemographics) to assess whether the model performs appropriately on the groups, including minorities. There is often a tension in this field as excluding demographic variables might hide structural inequalities rather than addressing them, which is something that should be carefully considered.

6.3 Implications for Stakeholders

The findings of this dissertation have implications that extend beyond the technical research community. They concern the daily practices of teachers, the learning experiences of students, the responsibilities of school boards, the innovation pathways of software vendors, and the regulatory perspectives of policy makers. Across all stakeholder groups, a central theme emerges: artificial intelligence (AI) in education should not aim to replace human expertise, but to enhance it through uncertainty-awareness, transparency, and efficiency. In this section, we discuss how each group might interpret and apply the results.

6.3.1 Teachers and Instructors

Teachers increasingly face pressure to personalize learning, identify at-risk students, and design fair assessments. First, this work calibrates teachers' **trust in predictions**. By showing that NNs can be "confidently wrong", it emphasizes the need for caution when adopting predictive tools. For instructors, this means that AI enhanced with UQ can act as a supportive tool: predictions are offered where the system is confident, while ambigu-

ous cases are explicitly flagged for human judgment. Second, it provides **assessment support** to teachers. The findings on QDE suggest that teachers can rely on data-efficient methods (active learning) to support exam preparation, reducing the time to evaluate large pools of potential exam questions. This is particularly valuable in a context where generative LLMs can rapidly produce candidate questions. Third, it improves the **grading fairness** in classrooms. By emphasizing the ordinal structure of assessment tasks, this work aligns algorithmic evaluation with teachers' natural grading practices. Models that recognize progressions such as "easy"–"medium"–"hard" or "A"–"F" provide outputs that are easier for instructors to interpret and to trust.

6.3.2 Students

For students, the most important outcomes concern fairness, transparency, and trust. First, this work **reduces harm from student misclassification**. When predictive models are overconfident in their errors, students risk being mislabeled, either as struggling when they are not, or as high-performing when they need support. By explicitly modeling uncertainty, the methods presented here reduce such risks and ensure that students are less likely to be incorrectly classified. Second, it provides **fairer evaluations** of students and learning material. Similarly, ordinal-aware approaches to grading and difficulty estimation reflect the incremental nature of learning, ensuring that automated judgments align more closely with students' actual performance levels. This reduces the risk of arbitrary or inconsistent outcomes and fosters trust in educational technologies. Third, it advocates for **responsible use of LLMs** in education. This work shows that current role-playing approaches with generative LLMs are not yet sufficiently reliable for high-stakes assessment. For students, this is a protective finding: it prevents premature adoption of technologies that might otherwise influence grading or feedback in ways that are inconsistent or biased.

6.3.3 School Boards and Administrators

School boards face the dual challenge of improving outcomes while using resources efficiently. First, this work guides **resource allocation** of school personnel. Predictive systems that incorporate UQ can signal when a model's output should not be trusted. When models defer uncertain cases to human experts, scarce support staff such as counselors, tutors, or advisors can focus their efforts on the students who most need attention. Second, transparent evaluation metrics such as balanced DRPS strengthen institu-

tional **accountability**. They provide school boards with principled ways to assess the reliability of AI tools before adoption, ensuring compliance with internal standards and external regulations. Third, as early-warning systems are deployed, it highlights the need to **carefully register** when and how **early-warning interventions** take place, possibly in collaboration with educational software vendors. Such records provide valuable supervision signals for assessing which interventions are most effective for different student groups, creating a feedback loop that can guide future decisions.

6.3.4 Educational Software Vendors

Vendors are under pressure to innovate quickly while maintaining trustworthiness. First, this work offers insights that support **product differentiation**. Incorporating UQ into predictive systems allows vendors to market their tools as not only powerful but also trustworthy. A learning analytics dashboard that explains when the AI is uncertain is more attractive to schools than a black box. Second, the findings highlight strategies that help **reduce costs**. Active learning provides a practical advantage for vendors: it reduces the annotation costs associated with building assessment systems. This lowers development costs and accelerates innovation cycles, making it feasible to extend ML systems to multiple subjects and educational contexts. Third, it promotes **user adoption** as ordinal-aware metrics and models produce outputs that align with teachers' intuitions. This makes these systems more interpretable and user-friendly, which increases the likelihood of long-term adoption.

6.3.5 Policy-Makers and Regulators

For policy-makers, the key concern is responsible integration of AI into education. First, this work aligns directly with principles of **risk mitigation** and responsible AI governance. The integration of human-in-the-loop design, through classification with rejection, offers a concrete mechanism for ensuring that AI systems do not override human expertise in high-stakes settings. Second, it increases **transparency** and allows justifying assessment outcomes. By explicitly addressing ordinality in assessment content, the research enables AI systems to be more aligned with human expectations. Third, it provides a foundation for pilot studies, which in turn guide an **evidence-based policy** on education. The limitations identified in the use of LLMs for student simulation caution against premature adoption of these tools in high-stakes testing or large-scale assessment. The findings

in this work advise a careful, phased integration of AI, rather than rapid deployment driven by hype.

6.4 Limitations and Future Research

An important limitation of the student performance prediction methodology in Chapter 2 is its assumption that no prior interventions have taken place and that the intervention has a uniform positive effect across all students. While the assumption about prior interventions holds for the dataset used in our study, it will become less realistic as early-warning systems gain wider adoption in educational practice. Moreover, not all students respond equally to additional support; for example, interventions directed at demotivated students may not yield the intended outcomes.

When data on prior interventions is available, *causal ML techniques* (Imbens & Rubin, 2015) offer a promising alternative. Instead of merely identifying students most likely to fail, causal approaches enable the identification of those most likely to benefit from support. This reframes the objective from risk detection to impact optimization and raises important ethical questions regarding how support should be allocated. Ultimately, such approaches compel instructors to balance effectiveness, fairness, and resource constraints in their intervention strategies. Deploying causal ML methods in real-life settings also calls for UQ, which is currently under-explored in literature.

Instructors also have multiple modes of intervention available—ranging from low-cost methods such as automated emails or instant messaging to high-effort actions like face-to-face consultations (Wong & Li, 2020). These vary in both cost and effectiveness. Methods at the intersection of causal ML and *cost-sensitive ML* can assist in selecting the optimal intervention for each student, balancing pedagogical effectiveness with practical resource limitations (Verbeke et al., 2023).

In Chapter 3, question difficulty is modeled using a small number of coarse labels. Future work could extend this to more fine-grained difficulty levels, as in the ARC dataset (Clark et al., 2018), which distinguishes between seven grade levels from 3 to 9. In such settings, the ordinal structure of the labels becomes more prominent, making the balanced DRPS metric and OrderedLogitNN model proposed in Chapter 4 especially relevant.

Furthermore, the OrderedLogitNN model naturally produces probability distributions over ordered classes, offering a straightforward way to quantify aleatoric uncertainty. When combined with methods such as MC Dropout or Deep Ensembles, it can also capture epistemic uncertainty—facilitating downstream tasks like classification with rejection and active

learning.

An interesting avenue of future work is to further investigate the model behavior of OrderedLogitNN proposed in Chapter 4, varying the weight initializations and learning rate multiplier. Future work could benchmark it on other modalities, such as image data, e.g., age estimation from facial images, a common ordinal regression task (Cao et al., 2020; Shi et al., 2023). It would also be valuable to assess how UQ from OrderedLogitNN performs in human-in-the-loop tasks such as classification with rejection and active learning. This would help assess its utility in scenarios where uncertainty plays a central role in downstream decision-making.

Furthermore, while this dissertation applies UQ on discriminative LLMs (either on tabular or text data), recent work has begun to explore *UQ for generative LLMs* (Shorinwa et al., 2025). This is crucial for identifying hallucinations and unreliable outputs, calibrating users' trust in generative models from Chapter 5. In the context of virtual pretesting, uncertain simulated responses could be filtered or rejected to improve the reliability of the difficulty estimation.

For QDE in general, a logical next step is its integration into *exercise recommender systems* for personalized learning. By matching students' skill levels with appropriate question difficulties, such systems can adaptively guide learners through tailored sequences of exercises. The research group is actively exploring contextual bandits (Langford & Zhang, 2007) for this purpose—an extension of the multi-armed bandit model by incorporating contextual information (e.g., student performance, question features) into the decision-making process.

Existing QDE research remains overly text-centric, with limited attention to multimodal assessment formats. In domains such as mathematics and sciences in general, exam questions frequently rely on diagrams, graphs, or other visual elements. Current research typically ignores these components. This limits the validity of ML-based systems, as real students process both textual and visual information when solving problems. Research into *vision-language models* for educational contexts would therefore bring ML systems closer to authentic student reasoning, enhancing both their accuracy and their relevance.

The methodologies developed for QDE also generalize to *distractor evaluation* in multiple-choice questions (Benedetto et al., 2025). Its importance has been highlighted in two recent Kaggle competitions (“Eedi - Mining Misconceptions in Mathematics” (King et al., 2024) and “MAP - Charting Student Math Misunderstandings” (King et al., 2025)). Distractor quality is a key determinant of overall question quality, e.g., if they are obviously wrong, they are never selected. While distractors are easy to

generate using LLMs, their quality is difficult to assess. As with QDE, ML approaches could serve as scalable approximations to human evaluation.

Although this dissertation focuses on challenges within the education domain, the broader implications extend to AI4SG. The challenges of generalization, data scarcity, and evaluation misalignment are not unique to learning analytics. As noted by Karamolegkou et al. (2025), similar issues arise in other social domains such as healthcare, environmental protection, and misinformation. The methodological contributions presented here—including UQ, ordinal modeling, active learning, and in-context learning for LLM-based role-playing—have the potential to support the development of more robust, fair, and responsible ML systems across these domains.

References

- Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., & Maximov, Y. (2025). Self-training: A survey. *Neurocomputing*, *616*, 128904.
- Benedetto, L., Taslimipoor, S., & Buttery, P. (2025). A survey on automated distractor evaluation in multiple-choice tasks, 55–69.
- Cao, W., Mirjalili, V., & Raschka, S. (2020). Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, *140*, 325–331. <https://doi.org/10.1016/j.patrec.2020.11.008>
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*. <https://doi.org/10.48550/arXiv.1803.05457>
- Croce, F., Rebuffi, S.-A., Shelhamer, E., & Goyal, S. (2023). Seasoning model soups for robustness to adversarial and natural distribution shifts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12313–12323.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
- Karamolegkou, A., Borah, A., Cho, E., Choudhury, S. R., Galletti, M., Ghosh, R., Gupta, P., Ignat, O., Kargupta, P., Kotonya, N., et al. (2025). Nlp for social good: A survey of challenges, opportunities, and responsible deployment. *arXiv preprint arXiv:2505.22327*. <https://doi.org/10.48550/arXiv.2505.22327>

- King, J., Burleigh, L., Woodhead, S., Kon, P., Baffour, P., Crossley, S., Reade, W., & Demkin, M. (2024). Eedi - mining misconceptions in mathematics [Kaggle].
- King, J., Smith, K., Burleigh, L., Crossley, S., Demkin, M., & Reade, W. (2025). Map - charting student math misunderstandings [Kaggle].
- Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459–9474.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.
- Shi, X., Cao, W., & Raschka, S. (2023). Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26(3), 941–955. <https://doi.org/10.1007/s10044-023-01181-9>
- Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2025). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.
- Verbeke, W., Olaya, D., Guerry, M.-A., & Van Belle, J. (2023). To do or not to do? cost-sensitive causal classification with individual treatment effect estimates. *European Journal of Operational Research*, 305(2), 838–852. <https://doi.org/https://doi.org/10.1016/j.ejor.2022.03.049>
- Vovk, V., Gammernan, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
- Wang, D., Shelhamer, E., Liu, S., Olshausen, B. A., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. <https://openreview.net/forum?id=uXl3bZLkr3c>

- Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8), 5362–5383.
- Wong, B. T.-m., & Li, K. C. (2020). A review of learning analytics intervention in higher education (2011–2018). *Journal of Computers in Education*, 7(1), 7–28.