Assessing LLMs Responses in the Field of Domestic Sustainability: an Exploratory Study

Mathyas Giudici
Politecnico di Milano
Milan, Italy
mathyas.giudici@polimi.it

Giulio Antonio Abbo
IDLab-AIRO
Ghent University – imec
Ghent, Belgium
giulioantonio.abbo@ugent.be

Ottavia Belotti
Politecnico di Milano
Milan, Italy
ottavia.belotti@mail.polimi.it

Alessio Braccini
Politecnico di Milano
Milan, Italy
alessio.braccini@mail.polimi.it

Francesco Dubini
Politecnico di Milano
Milan, Italy
franceso.dubini@mail.polimi.it

Riccardo Andrea Izzo
Politecnico di Milano
Milan, Italy
riccardo.izzo@mail.polimi.it

Pietro Crovari
Politecnico di Milano
Milan, Italy
pietro.crovari@polimi.it

Franca Garzotto
Politecnico di Milano
Milan, Italy
franca.garzotto@polimi.it

***Abstract*—In the next years, we must challenge climate change, and the urgency of adopting a more sustainable lifestyle has increased. Conversational Agents, such as Smart home Personal Assistants, have shown promise in fostering sustainable behaviors in domestic environments. However, traditional conversations with rule-based approaches in such agents face challenges in addressing users' questions in complex domains like sustainability. Large Language Models (LLMs) are a promising tool to overcome these limitations of their capability to answer open-domain questions. The final objective of this work is to compare the generative capabilities of four large language models in ecological sustainability to determine the most suitable LLM to be embedded into home assistants and create a hybrid model of conversational agent for environmental sustainability. We performed two evaluations. In the former, we constructed a set of trustable sources on the topic and analyzed the extent to which the themes covered in the text generated by the models appeared in it. The results do not show a statistical difference between the outputs of the candidate models, while qualitative analysis determined that ChatGPT, at the moment, is the optimal solution. In the second evaluation, we tested the responses generated by ChatGPT on a corpus of 167 questions from a sample of 75 people. Responses evaluation was performed by a team of experts (N=5) on fluency, coherency, consistency, accuracy, and reasoning. The results suggest that ChatGPT for generic questions on sustainability is quite reliable.**

***Index Terms*—Conversational Agent, Sustainability, LLM, Rule-based CA**

I. Introduction

It has been several years since the climate crisis first captured global attention [1]. The urgency to adopt a more sustainable lifestyle has risen, with the objective of containing both present and future damages to the environment [2]. However, understanding which behaviors contribute to achieving this objective – or even discerning the detrimental impact of certain habits – is often difficult [3]. The two main reasons are the complexity of the topic [4], [5] and the confusion caused by widespread misinformation [6], [7]. In 2010, [8] explored the role that human-computer interaction can have in shifting people towards more environmentally sustainable behaviors.

Among the different digital technologies nowadays available, Conversational Agents (CA) – i.e., technologies that interact with users through natural language [9] – are a promising tool. Their ability to interact in plain language allows users to focus entirely on the interaction goal, i.e., addressing sustainability issues, rather than on the interaction itself. Conversational Agents are deployed as stand-alone applications or embedded in larger systems. Among those, there are Smart home Personal Assistants (SPAs), IoT devices that provide a conversational interface to a domotic IoT environment [10]. These systems have been revealed to be valid solutions to foster sustainability behaviors in domestic environments [11], [12] for two main reasons: they are widely adopted and they can provide actionable feedback, since they are already connected with smart environments, and can measure and monitor users habits, for example in terms of energy consumption.

Nowadays, most Conversational Agents embedded in SPAs operate based on a predefined set of rules [13], [14], which means that developers – or conversation designers – need to anticipate and account for every potential scenario and user request. This approach is designed to ensure that interactions with the CA are predictable and follow a predetermined flow [15]. However, when it comes to addressing users' questions, particularly in a multifaceted and diverse domain like sustainability, this rule-based approach presents challenges, as the CA may struggle to provide satisfactory responses to unanticipated questions.

To overcome these limitations, newer approaches, such as generative conversational agents, are being developed [16]. These models leverage large datasets and advanced natural language processing techniques to understand and respond to user queries more dynamically. In particular, *Large Language Models* (LLMs) [17], neural networks that process text input and produce a textual output, can be used to expand the virtual assistants' capabilities thanks to their ability to answer open-domain questions [18]. However, generative models might not match the high standards in terms of content and scientific

accuracy required when the objective is informative and factual [19]. This weakness derives from biases in training data, limited access to real-time information, and the inability to verify facts. These challenges can lead to generating responses that are inaccurate, outdated, or misleading.

To address these issues, additional measures such as fact-checking, human oversight, and expert validation are necessary. However, implementing these steps can often be cost-prohibitive and may not be feasible or practical in many situations, presenting a challenge in achieving consistently accurate and reliable responses. For these reasons, we want to understand whether LLMs are able to deliver precise and scientifically accurate answers on the sustainability topic, with a particular focus on the household domain. We develop our research in two phases. First, we present a comparison between different LLMs, testing which of the most prominent LLMs are best suited to deliver factual information on topic-related questions. Second, we gathered questions on sustainability with a crowdsourced questionnaire that involved 73 people. The corpus created is then analyzed and used to evaluate the LLM obtained from the first part of the study, asking energy and sustainability five experts to evaluate the generated replies on fluency, coherence, consistency, accuracy, and argument logic. The results suggest that ChatGPT for generic questions on sustainability is quite reliable. However. experts reported low accuracy in the content delivered in the responses.

This work contributes to the application research for LLMs in the field of environmental sustainability while ensuring scientifically accurate information and human-like responses. We believe that our results have proven that the way of hybridization of rule-based and generative-based smart assistants can lead to upgraded tools able to provide effective support to householders in learning and adopting more sustainable lifestyles.

II. State of the Art

A. Large Language Models (LLMs)

A Large Language Model (LLM) is a neural network model specifically trained, usually on a billion parameters, to understand, summarize and generate text-based content [20]. The applications for LLMs are endless, some notable examples being summarization, as presented by [21], and Next Sequence Prediction (NSP) [22]. However, the most pertinent usage for our purposes is employing LLMs as Conversational Agents.

OpenAI's *GPT-2*, released in 2019, can be considered the pioneer in LLM, and it was trained on 1.5 billion parameters [23]. With LLMs in the spotlight due to their recent impressive performance, big corporations decided to make a move and publish their own versions of LLMs. According to [24], some of the most notable ones nowadays are Google's *Bard* [25], Microsoft's *Bing AI* [26] and OpenAI's *ChatGPT/GPT-3.5* [27].

As explained by [28], the large size of modern language models has rendered traditional weight updates impractical. The unavailability and resource constraints make full model tuning unfeasible for many applications. As a result, the field of *prompting* emerged, exploring methods to leverage LLMs inputs for influencing the output. This method leverages zero-shot learning, a technique by which you can reach state-of-the-art performances by presenting examples of a given task to a pre-trained model without fine-tuning it [29]. In fact, in the case of prompting, the examples are fed via some text (a prompt) that steers the LLM toward giving the right answer. It is also worth noting how avoiding fine-tuning the models is an environmentally sustainable choice since the heavy GPU training oftentimes results in high carbon dioxide emissions, as emerged from the work of [30].

To the best of our knowledge, within the sustainability field, AI has been widely implemented during the past years [31], but no significant effort was made to implement LLMs as a channel to spread information about the environment and sustainable behavior.

B. Conversational Agents for Question Answering

Conversational Agents (CA) take advantage of natural language processing techniques to engage users in text-based information-seeking and task-oriented dialogues for a multitude of applications [32]. For example, they are integrated into physical devices (such as Alexa and Google Home) and available in many contexts of everyday life, used in phones (like Siri, the Apple virtual assistant), cars, and online consumer assistance [33]. They also find extensive use in applications such as question answering, leading to the development of Conversational Question Answering (CQA) systems [34]. CQA systems aim to comprehend given context and manage single-turn or multi-turn question answering to satisfy a user's information needs. Finally, according to previous studies [35], [36], conversational technologies are a promising interaction paradigm to persuade people towards more sustainable behaviors, previously successfully used in different settings.

ELIZA [37] represents a significant historical reference in the field of conversational agents. As a rule-based model, ELIZA used predefined rules and lexicons for language generation and comprehension, exploiting the concept of pattern matching [38] that set up the stage for advanced models.

Although rule-based CAs have been a milestone in this mode of interaction in the past, they had significant limitations, such as a lack of adaptability and scalability of conversations [39].

In the 1980s, there was a shift from rule-based systems to data-driven approaches [40] powered by probability distributions over sequences of words [41]. Finally, in the 2000s, neural networks and the Transformer model were introduced for the first time by [42]. These technologies, combined with the advancements in LLMs (previously described in Section II-A), led to the emergence of modern Conversational Agents such as ChatGPT.

In the environmental sustainability field, rule-based approaches were used for delivering energy feedback [43], [44], suggesting sustainable mobility [45], or reducing food

waste [46]. Instead, [47] is an example using data-driven systems to suggest recipes with leftover foods.

III. Extending Home assistant with LLM capabilities

As explained in the previous sections, prompting is one of the primary tools in tuning large language models. However, such a technique has some limitations [48].

One of the most current challenging limitations is making the prompt change depending on some real-time variable, which also implies adjusting the logic of such models to incorporate the variables. An approach to overcome that issue is proposed in the Socratic models by [49]. All the data sensed by the external environment are fed inside the prompt and the LLM replies accordingly. However, such a solution relies on the input length of the generative model, becoming impracticable if the contextual data are too long.

Another approach is the use of hybrid models; they handle requests related to real-time data retrieval through a rule-based strategy while managing all the other inquiries with a generative approach. Although considerable effort has been made in the past literature to present different approaches of hybrid model mixing generative and retrieval techniques [50], the research panorama for a hybrid model for environmental sustainability is still scant.

IV. Exploratory Evaluation: LLM Selection

We carried out an explorative evaluation, represented in Figure 1, aiming to find out which generative models have the potential to be integrated into conversational agents and talk about environmental sustainability in a domestic context.

To achieve this objective, we set up a study to investigate two aspects. First, how domestic sustainability information is included in a text generated by an LLM and which topics are touched, in comparison to those covered by a set of reliable sources, as shown in Figure 2. Second, the value of the variability of topics delivered by the same LLM across multiple runs.

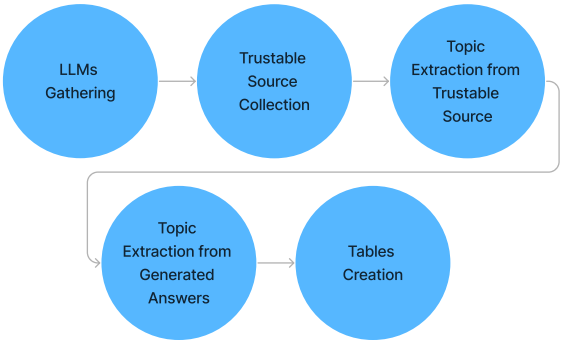

Fig. 1. Steps of the Exploratory Evaluation.

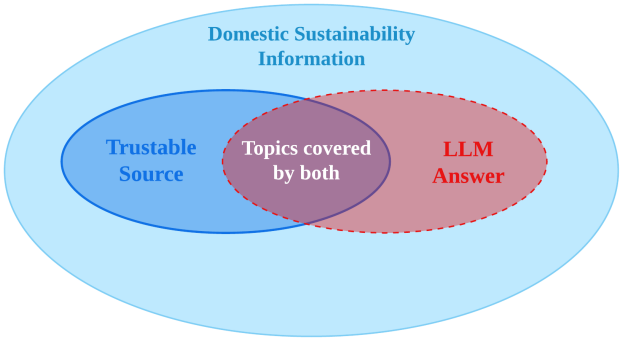

Fig. 2. Graphical Representation of Domestic Sustainability Information Domain.

We considered for this evaluation the most prominent LLMs [51] at the time of writing (i.e., end of June 2023), those considered state-of-the-art and possibly accessible via API. The LLMs evaluated in this stage were *ChatGPT*, *BingAI*, *Bard*, and *OpenAssistant* (LLAMA version).

We built a set of ten questions on the topic of sustainability (reported in Appendix A[1]). To mitigate the introduction of human bias during this process, we relied on news articles (from newspapers, scientific periodicals, reliable scientific blogs, etc.) considered a credible source of information by numerous people [52]. We considered the first 10 articles on the topic, published before 2021 (otherwise, the topics covered might not be part of the corpora on which most of the models are built), and that had the title or the main line formulated as questions. The body of each article represents the trustable answer to each question.

From these 10 article bodies, we extracted the topics covered. We asked three subjects who did not have previous experience in the field to list the topics covered by each article, i.e., an argument that is discussed in sufficient detail to comprehend its meaning. Based on this data, content topics are assigned, applying a majority voting technique.

We sampled the four candidate LLMs, asking the questions extracted from the trustable source, using a zero-shot promoting technique. We asked the same question 8 times to each model to allow for variance in the answers and investigate whether different topics appeared. To ensure that there were no correlations between the different generations, a new instance of each model was created after each response. We then extracted the topics from the answers, using the same methodology adopted for the trustable source.

Finally, we created a table to match the topics identified in the answers generated by the LLMs (and their occurrence) and the ones extracted from the trustable source. An example reporting the topic extraction of the first question is reported in Table I.

[1]https://doi.org/10.5281/zenodo.10012799

TABLE I
EXAMPLE OF TOPIC EXTRACTION AND OCCURRENCES FOR QUESTION 1: *Where does all the carbon we release go?*

Topic	Trustable source	ChatGPT	BingAI	Bard	LLama
Original source	✓	7	0	7	5
Atmosphere cycle	✓	6	5	6	6
Human carbon production	✓	7	0	2	7
Solutions to reduce carbon	✓	3	6	2	3
Global warming	✓	5	0	0	5
Why carbon is necessary		1	2	4	0

A. Results and Discussion on the Exploratory Evaluation

Since we mapped the presence of topics in the 8 responses generated by the LLMs to each individual question. The result reports that, on average, a trustable source topic is covered 0.850 by *ChatGPT* (M=0.850, SD=0.285). For *BingAI*, the average is 0.655 (SD=0.273), while *Bard* has a mean of 0.728 (SD=0.309). Finally, *LLama* stands with a mean of 0.622 (SD=0.240).

TABLE II
DESCRIPTIVE RESULTS OF ALIGNMENT OF TOPICS BETWEEN TRUSTABLE SOURCE AND LLM

			Percentiles		
	Mean	SD	25th	50th	75th
ChatGPT	0.850	0.285	0.800	0.857	1.000
BingAI	0.655	0.273	0.425	0.686	0.843
Bard	0.728	0.309	0.600	0.657	0.964
Llama	0.622	0.240	0.450	0.675	0.779

In addition, we run a comparative analysis to evaluate the alignment between the topics presented in the trustable sources and those reflected in the LLMs' answers, examining the variance between multiple completions from the same model.

As shown in Table III, every topic appears on average in 4.41 out of 8 questions generated by *ChatGPT* (M=4.41, SD=2.39) to the same question. For *BingAI*, every topic appears on average 4.01 times (SD=1.96), while on *Bard*, they are inserted in 4.55 answers (SD=1.99). Finally, *LLama* repeats the same topics around 4.40 (SD=1.97) on the 8 answers.

TABLE III
DESCRIPTIVE RESULTS OF TOPICS EXTRACTED BY GENERATED RESPONSES

	Mean	SD	Min	Max
ChatGPT	4.41	2.39	1	8
BingAI	4.01	1.96	1	7
Bard	4.55	1.99	1	7
LLama	4.40	1.97	1	7

We run Repeated Measures ANOVA to assess the variability in the answers produced in terms of the number of topics covered (among the repeated runs) in answers to the same question by the four LLMs (all the results are reported in Table IV). There is no statistically significant difference in the number of topic generation ($F(9,3)=1.386$, $p=0.250$), also running the Friedman non-parametric test ($X^2(3)=5.75$, $p=0.124$).

Given the results of this preliminary evaluation, there is no LLM model that statistically outperforms the others in terms of the number of topics covered in the generated answers. In the same way, we do not have statistical evidence of more recurrent topics in the generated answers.

For this reason, for the scope of our study described in the following section, we select an LLM model on qualitative observations. During the topic extraction, we can report that qualitatively the impression is that *ChatGPT* and *Bard* are very verbose in their responses. On the other hand, *BingAI* is very brief in its responses but equally perceived as accurate in its alignment with the trustable source.

Since the underlying motivation that led to this study was to integrate such a question-answering system into a SPA, we decided to select *ChatGPT*, since – at the time of the study – was the only system offering an API interface, therefore natively supporting integration.

V. EMPIRICAL STUDY: LLM ANSWERS EVALUATION

In this second study, represented in Figure 3, the objective is to analyze and evaluate the responses that the LLM chosen with the previous exploratory evaluation could provide to a user. To achieve this, we want to answer the following research question: *What is the performance of responses generated by an LLM from a crowdsourced dataset of questions on the topic of environmental sustainability?*

Fig. 3. Steps of the Empirical Study.

A. Methodology

We will assess the responses generated by the LLM using a set of metrics grounded on previous works on a similar topic [53], [54]. For each dimension, we gathered five scores (on a 5-item Likert scale).

- *Fluency* measures how well the response is written in natural language, without syntactic errors or awkward phrasing.
- *Coherency* measures how a response is free of logical contradictions.

TABLE IV
ANOVA RESULTS OF TOPICS EXTRACTED BY GENERATED RESPONSES

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p
LLM	11.2	3	3.72	1.386	0.250
LLM * Question	71.1	27	2.63	0.981	0.499
Residual	362.5	135	2.69		

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p
Question	38.4	9	4.27	0.461	0.893
Residual	416.9	45	9.27		

Note. Type 3 Sums of Squares

- *Consistency* measures how much the topic of the response aligns with the topic in the question.
- *Accuracy* measures how much the response is factual and accurate with respect to the topic.
- *Argumentation* measures how much the response is well explained, without redundancy or lacking in common sense.

Having defined the evaluation metrics, we proceeded to acquire a set of questions on the topic of sustainability to assess the performance of the chosen LLM (i.e., ChatGPT). To this extent, we created a questionnaire asking participants to write 3 questions they would ask an expert in the sustainability field. Participants belonged to close contacts from the personal community, colleagues, or university students (the latter with a predominantly scientific background), and they were all sensitive to environmental sustainability issues (without being able to consider them experts).

We collected the results into a CSV file, analyzed them with the Sklearn K-means Clustering algorithm and using an embedding created with BERT, and extracted the main categories into which they could be divided. Then, the clusters were labeled.

We selected a random question from each category identified in the previous step, and we generated a response with the LLM. The prompt contained the sentence *"Explain in 50 words:"* followed by the text of the question.

Finally, to assess the performance of the generated responses, we asked a group of 5 experts specialized in the field of sustainability (currently working in major energy companies in Italy) to evaluate the generated answers through a specific questionnaire (see Appendix B and C[2]), following the metrics presented above.

B. Results and Discussion

In total, 75 respondents completed the questionnaire and provided in total 225 partially overlapping questions. After having discarded low-quality contributions (i.e., questions not related to the sustainability field), and removed the duplicates, we were left with 167 unique questions.

We classified the questions collected into 7 distinct categories, as shown in Figure 4, where some points are mixed with others due to the t-SNE dimensionality reduction used to visualize the graph in two dimensions.

The clusters were labeled based on the topics and the labels are here reported together with the number of unique answers associated: *Reducing plastic use* (6), *Concerns about the environment and related actions* (5), *Environmental policies* (44), *Greenwashing and "green" solutions* (9), *Eco-sustainability and sustainable behaviors* (27), *Energy and environmental impact* (6), and *Environmental sustainability* (70).

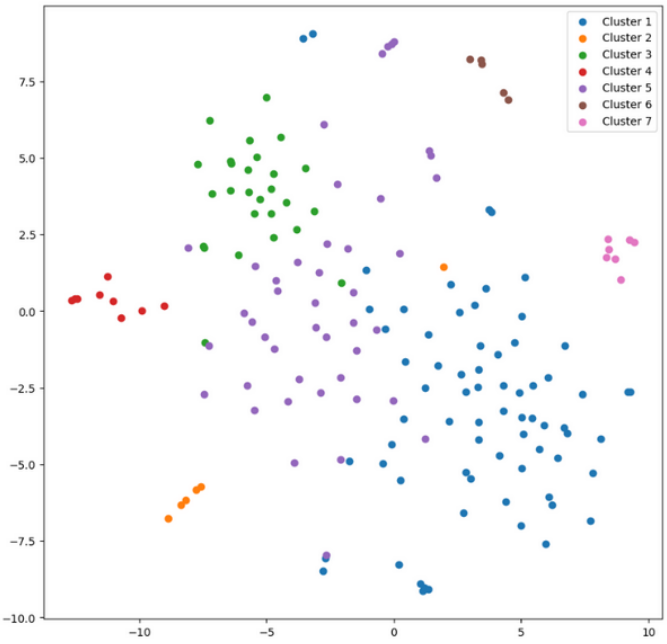

Fig. 4. Sklearn K-mean Clustering Analysis (points of different colors overlap each other due to the dimensional reduction caused by t-SNE algorithm)

The questions randomly selected for the evaluation and the corresponding generated answers are reported in the supporting material. In Table V, we report the results of the evaluation of these responses by a group of (N=5) experts in the field.

Experts reported fluency in the generated answers with an average of 4.49 (SD=0.612). The coherency mean is 4.46 (SD=0.611), while consistency has M=4.11 and SD=1.022, and accuracy has M=3.71 and SD=1.045. Finally, argumentation has M=3.80 and SD=0.994.

[2]https://doi.org/10.5281/zenodo.10012799

TABLE V
DESCRIPTIVE RESULTS OF SCORES ASSIGNED TO CHATGPT

	Mean	SD	Min	Max
Fluency	4.49	0.612	3.00	5.00
Coherency	4.46	0.611	3.00	5.00
Consistency	4.11	1.022	2.00	5.00
Accuracy	3.71	1.045	2.00	5.00
Arguments Logic	3.80	0.994	2.00	5.00

The experts agreed that *ChatGPT* can answer questions about domestic sustainability in a fluent and coherent way. However, they also reported a low accuracy in the content of the generated responses. Our result is contrarily to the performance reported in other domains like [55].

We believe that a fine-tuning phase might mitigate such a problem in the sustainability field before the deployment of the system or by extending the prompt with additional context information.

In addition, qualitative insights provided by experts highlighted that the lower accuracy in the answers seems to correspond to questions provided by users that are difficult to interpret, even to human respondents. In the future, we will to further understand the relation between the quality of the questions and the obtained output, trying to elicit strategies to minimize its influence in the generation of the responses.

VI. CONCLUSION

In an exploratory evaluation, we compared the generative capabilities of four large language models in the field of ecological sustainability, with the objective of determining the most suitable to be embedded in a conversational agent in the home environment. The models considered are *ChatGPT*, *BingAI*, *Bard*, and *LLAMA*.

We constructed a set of trustable sources on the topic and analyzed the extent to which the themes covered in the text generated by the models appeared in it. We sampled each model multiple times and performed an ANOVA comparison. The data gathered was not sufficient to highlight a statistical difference between the outputs of the candidate models. However, the results differed in terms of verbosity and we determined that *ChatGPT*, at the moment, is the optimal solution.

To test the responses generated by *ChatGPT*, we built a corpus of 167 questions on the topic of sustainability from a group of 75 participants. We then used *ChatGPT* to produce candidate answers to these questions and evaluated them in terms of fluency, coherency, consistency, accuracy, and argumentation. The results confirm that *ChatGPT* as-is can answer very general questions and is quite reliable. However, the low accuracy reported by experts in the questionnaire points out, that for some specific questions, there is the need to add prompting or use fine-tuning techniques. Future work will allow us to test all the questions in each cluster and then correlate them with the various scores. In addition, we will expand the number of experts in evaluating the responses generated to user questions.

ACKNOWLEDGMENT

This work is partially supported by the Italian Ministry of University and Research under the PONRI program (Programma Operativo Nazionale "Ricerca e Innovazione") 2014-2020 and by the European Commission under Horizon Europe program - VALAWAI project (grant agreement number 101070930).

REFERENCES

[1] D. Archer and S. Rahmstorf, *The climate crisis: An introductory guide to climate change*. Cambridge University Press, 2010.

[2] R. Pierrehumbert, "There is no plan b for dealing with the climate crisis," *Bulletin of the Atomic Scientists*, vol. 75, no. 5, pp. 215–221, 2019.

[3] L. Carrete, R. Castaño, R. Felix, E. Centeno, and E. González, "Green consumer behavior in an emerging economy: confusion, credibility, and compatibility," *Journal of Consumer Marketing*, vol. 29, no. 7, pp. 470–481, Jan 2012. [Online]. Available: https://doi.org/10.1108/07363761211274983

[4] M. T. Boykoff, *Who speaks for the climate?: Making sense of media reporting on climate change*. Cambridge University Press, 2011.

[5] J. D. Sterman, "Sustaining sustainability: creating a systems science in a fragmented academy and polarized world," *Sustainability science: The emerging paradigm and the urban environment*, pp. 21–58, 2012.

[6] P. Parks, "Is climate change a crisis–and who says so? an analysis of climate characterization in major us news media," *Environmental Communication*, vol. 14, no. 1, pp. 82–96, 2020.

[7] F. Biondo, G. La Rocca, and V. Viviana Trapani, "Information disorder: Learning to recognize fake news," 2022.

[8] C. DiSalvo, P. Sengers, and H. Brynjarsdóttir, "Mapping the landscape of sustainable hci," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010.

[9] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A survey on conversational agents/chatbots classification and design techniques," in *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 2019, pp. 946–956.

[10] J. Santos, J. J. Rodrigues, J. Casal, K. Saleem, and V. Denisov, "Intelligent personal assistants based on internet of things approaches," *IEEE Systems Journal*, vol. 12, no. 2, pp. 1793–1802, 2016.

[11] N. Shevchuk and H. Oinas-Kukkonen, "Exploring green information systems and technologies as persuasive systems: A systematic review of applications in published research," 12 2016.

[12] B. J. Fogg, "Persuasive technology: Using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, dec 2002. [Online]. Available: https://doi.org/10.1145/764008.763957

[13] A. J. Obaid, "Assessment of smart home assistants as an iot," *International Journal of Computations, Information and Manufacturing (IJCIM)*, vol. 1, no. 1, 2021.

[14] N. Abdi, K. M. Ramokapane, and J. M. Such, "More than smart speakers: Security and privacy perceptions of smart home personal assistants." in *SOUPS@ USENIX Security Symposium*, 2019.

[15] K. P. Jadhav and S. A. Thorat, "Towards designing conversational agent systems," in *Computing in Engineering and Technology: Proceedings of ICCET 2019*. Springer, 2020, pp. 533–542.

[16] K. Ramesh, S. Ravishankaran, A. Joshi, and K. Chandrasekaran, "A survey of design techniques for conversational agents," in *Information, Communication and Computing Technology: Second International Conference, ICICCT 2017, New Delhi, India, May 13, 2017, Revised Selected Papers*. Springer, 2017, pp. 336–350.

[17] M. Zaib, Q. Z. Sheng, and W. Emma Zhang, "A short survey of pre-trained language models for conversational ai-a new age in nlp," in *Proceedings of the Australasian computer science week multiconference*, 2020, pp. 1–4.

[18] G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, and K. Lee, "Prompted llms as chatbot modules for long open-domain conversation," *arXiv preprint arXiv:2305.04533*, 2023.

[19] P. Pichappan, M. Krishnamurthy, and P. Vijayakumar, "Analysis of chatgpt as a question-answering tool," *Journal of Digital Information Management*, vol. 21, no. 2, p. 51, 2023.

[20] A. Lee, "What Are Large Language Models and Why Are They Important? — NVIDIA Blog — blogs.nvidia.com," https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/, 2023, [Accessed 07-Jun-2023].

[21] A. Hoang, A. Bosselut, A. Celikyilmaz, and Y. Choi, "Efficient adaptation of pretrained transformers for abstractive summarization," *CoRR*, vol. abs/1906.00138, 2019. [Online]. Available: http://arxiv.org/abs/1906.00138

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[23] P. Budzianowski and I. Vulić, "Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems," *arXiv preprint arXiv:1907.05774*, 2019.

[24] J. Rudolph, S. Tan, and S. Tan, "War of the chatbots: Bard, bing chat, chatgpt, ernie and beyond. the new ai gold rush and its impact on higher education," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, 2023.

[25] S. Pichai, "An important next step on our ai journey," Feb 2023. [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates/

[26] Y. Mehdi, "Reinventing search with a new ai-powered microsoft bing and edge, your copilot for the web," May 2023. [Online]. Available: https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[27] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of chatgpt/gpt-4 research and perspective towards the future of large language models," 2023.

[28] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023. [Online]. Available: https://doi.org/10.1145/3560815

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[30] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," 2019.

[31] S. S. Biswas, "Potential use of chat gpt in global warming," *Annals of Biomedical Engineering*, vol. 51, no. 6, pp. 1126–1127, Jun 2023. [Online]. Available: https://doi.org/10.1007/s10439-023-03171-8

[32] J. Lester, K. Branting, and B. Mott, "Conversational agents," *The practical handbook of internet computing*, pp. 220–240, 2004.

[33] R. Bavaresco, D. Silveira, E. Reis, J. Barbosa, R. Righi, C. Costa, R. Antunes, M. Gomes, C. Gatti, M. Vanzin *et al.*, "Conversational agents in business: A systematic literature review and future research directions," *Computer Science Review*, vol. 36, p. 100239, 2020.

[34] M. Zaib, W. E. Zhang, Q. Z. Sheng, A. Mahmood, and Y. Zhang, "Conversational question answering: a survey," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3151–3195, Dec 2022. [Online]. Available: https://doi.org/10.1007/s10115-022-01744-y

[35] M. Giudici, P. Crovari, and F. Garzotto, "Candy: A framework to design conversational agents for domestic sustainability," in *4th Conference on Conversational User Interfaces*, ser. CUI 2022. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3543829.3544515

[36] L. Å. E. J. Hansson, T. Cerratto Pargman, and D. S. Pargman, "A decade of sustainable hci: Connecting shci to the sustainable development goals," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–19.

[37] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[38] H.-y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 10–26, Jan 2018. [Online]. Available: https://doi.org/10.1631/FITEE.1700826

[39] S. A. Thorat and V. Jadhav, "A review on implementation issues of rule-based chatbot systems," in *Proceedings of the international conference on innovative computing & communications (ICICC)*, 2020.

[40] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[41] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[43] U. Gnewuch, S. Morana, C. Heckmann, and A. Maedche, "Designing conversational agents for energy feedback," in *International Conference on Design Science Research in Information Systems and Technology*. Springer, 2018.

[44] M. Giudici, P. Crovari, and F. Garzotto, "Leafy: Enhancing home energy efficiency through gamified experience with a conversational smart mirror," in *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, ser. GoodIT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 128–134. [Online]. Available: https://doi.org/10.1145/3582515.3609526

[45] S. Diederich, S. Lichtenberg, A. B. Brendel, and S. Trang, "Promoting sustainable mobility beliefs with persuasive and anthropomorphic design: Insights from an experiment with a conversational agent," 2019.

[46] N. M. Cacanindin, "Greening Food Consumption Using Chatbots as Behavioral Change Agent," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 01-Special Issue, pp. 204–211, Feb. 2020.

[47] M. Gunawardane, H. Pushpakumara, E. Navarathne, S. Lokuliyana, K. Kelaniyage, and N. Gamage, "Zero food waste: Food wastage sustaining mobile application," in *2019 International Conference on Advancements in Computing (ICAC)*. IEEE, 2019, pp. 129–132.

[48] R. Cohen, M. Hamri, M. Geva, and A. Globerson, "Lm vs lm: Detecting factual errors via cross examination," *arXiv preprint arXiv:2305.13281*, 2023.

[49] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, "Socratic models: Composing zero-shot multimodal reasoning with language," 2022.

[50] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, and J. Liu, "A hybrid retrieval-generation neural conversation model," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1341–1350. [Online]. Available: https://doi.org/10.1145/3357384.3357881

[51] M. S. Rahaman, M. Ahsan, N. Anjum, M. M. Rahman, and M. N. Rahman, "The ai race is on! google's bard and openai's chatgpt head to head: an opinion article," *Mizanur and Rahman, Md Nafizur, The AI Race is on*, 2023.

[52] R. A. Abdulla, B. Garrison, M. Salwen, P. Driscoll, and D. Casey, "The credibility of newspapers, television news, and online news," in *Education in Journalism Annual Convention, Florida USA*, 2002.

[53] E. Reiter and A. Belz, "An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 12 2009. [Online]. Available: https://doi.org/10.1162/coli.2009.35.4.35405

[54] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, and X. Liu, "Ai vs. human – differentiation analysis of scientific content generation," 2023.

[55] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, "Orca: Progressive learning from complex explanation traces of gpt-4," 2023.