On LLM Wizards: Identifying Large Language Models' Behaviors for Wizard of Oz Experiments

Jingchao Fang*
Toyota Research Institute
USA
jcfang@ucdavis.edu

Nikos Arechiga
Toyota Research Institute
USA
nikos.arechiga@tri.global

Keiichi Namikoshi
Toyota Research Institute
USA
keiichi.namikoshi@tri.global

Nayeli Bravo
Toyota Research Institute
USA
nayeli.bravo@tri.global

Candice Hogan
Toyota Research Institute
USA
candice.hogan@tri.global

David A. Shamma
Toyota Research Institute
USA
ayman.shamma@tri.global

Abstract

The Wizard of Oz (WoZ) method is a widely adopted research approach where a human Wizard "role-plays" a not readily available technology and interacts with participants to elicit user behaviors and probe the design space. With the growing ability for modern large language models (LLMs) to role-play, one can apply LLMs as Wizards in WoZ experiments with better scalability and lower cost than the traditional approach. However, methodological guidance on responsibly applying LLMs in WoZ experiments and a systematic evaluation of LLMs' role-playing ability are lacking. Through two LLM-powered WoZ studies, we take the first step towards identifying an experiment lifecycle for researchers to safely integrate LLMs into WoZ experiments and interpret data generated from settings that involve Wizards role-played by LLMs. We also contribute a heuristic-based evaluation framework that allows the estimation of LLMs' role-playing ability in WoZ experiments and reveals LLMs' behavior patterns at scale.

CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**.

Keywords

Wizard of Oz, large language model, synthetic data, persuasive conversation, methods, WoZ, LLM

ACM Reference Format:
Jingchao Fang, Nikos Arechiga, Keiichi Namikoshi, Nayeli Bravo, Candice Hogan, and David A. Shamma. 2024. On LLM Wizards: Identifying Large Language Models' Behaviors for Wizard of Oz Experiments. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3652988.3673967

*Author is currently at UC Davis.

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673967

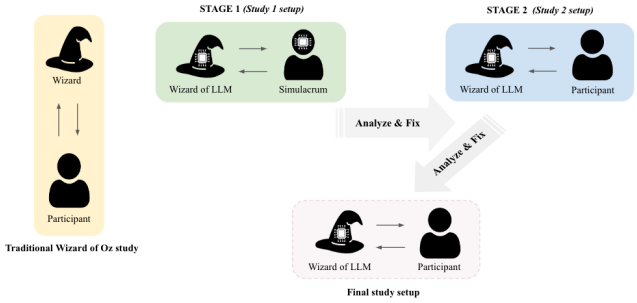


Figure 1: An overview of our proposed experiment lifecycle compared to traditional Wizard of Oz experiments. We ask GPT-4 empowered agents to play the role of "Wizards" in conversation-based Wizard of Oz experiments. The agents talk to either Simulacrums powered by GPT-4 (in Study 1) or Human Participants (in Study 2).

1 Introduction

People often have online conversations with individuals who possess specific information and expertise for help or facilitation; chatbots, deployed as conversational agents [21, 35, 58, 59, 71], offer the advantages of scalability and cost-effectiveness in these tasks. Consider implementing a chat agent to conduct persuasive conversations for social good (e.g., encouraging an environmentally friendly lifestyle). Developing the agent as an experimental device creates several hurdles. Training and fine-tuning a model requires an eco-friendly corpus of training data to acquire accurate domain knowledge and reduce the rate of producing faulty or harmful messages. This training must include data representing individuals with diverse backgrounds as climate-related persuasions require knowing a person's values to avoid backlash [23]. In addition, multiple stages of the user-centered design process are inherently iterative and require rounds of user participation [50], which challenges rapid ideation and prototyping. Ultimately, a considerable amount of resources (e.g., training data, computing power, labor) is needed before the chat agent is polished for early user testing. Thus, gauging target users' attitudes and interactions with the agent *before* putting in much development effort is usually desired.

The Wizard of Oz (WoZ) method [19, 30, 33] and its Oz of Wizard variant [57] could be helpful for this purpose. Both methods are designed to overcome experimentation obstacles by simulating automatic systems or humans when testing ideas with them is expensive or infeasible. In our persuasive chatbot example, we can set up a WoZ experiment where a human Wizard (experimenter) role-plays the to-be-developed technology and talks to participants. The experiment can elicit data revealing users' behaviors and attitudes when interacting with an envisioned technology before it is implemented, reducing the cost of design and development iterations. Yet, scaling up WoZ is challenging due to the required human labor for role-playing.

Closely parallel with the "role-playing" in the WoZ method, recent studies propose leveraging large language models (LLMs) to "role-play" and simulate human-to-human or human-to-agent chats and generate synthetic data with low cost [36, 53]. The advancement of LLMs points to the potential of harnessing LLMs' speedy generation ability to role-play Wizards and scale up WoZ experiments. An overarching question that needs to be addressed is whether we can reliably use LLMs to elicit data that can be translated into design and development insights as a human Wizard would do in traditional WoZ.

In this paper, we take the first step towards exploring the feasibility of applying LLMs in conversational WoZ experiments. **We present an experiment lifecycle (Figure 1) for safely piloting and integrating LLMs into WoZ experiments where GPT-4 empowered agents, instead of humans, role-play as Wizards at scale.** The goal of the LLM Wizards is the *elicitation* of users' reactions to an envisioned technology being simulated in WoZ experiments (e.g., a specialized chatbot conducting persuasive conversations for social good), which provides design and development insights, rather than becoming the envisioned technology itself. The lifecycle is demonstrated via two studies, where GPT-4 agents act as "Wizards" (named as Wizard of LLMs, or WoLs) in WoZ experiments to talk to Simulacrums (also GPT-4 agents) and Participants (humans). This WoZ process generates insights guiding the development of new tools by: (1) collecting data that unveils how users engage with the to-be-invented tools on a large scale, and (2) understanding design spaces and opportunities for improvement for the envisioned tools, based on observed limitations LLM Wizards.

Following traditional experimentation models (e.g., original WoZ methodology [31], *many are called / refine / few are called* framework [16]), the experiment lifecycle starts with a coarse, cheap, and large-scale WoLs-to-Simulacrums setting (Stage 1). While LLMs' role-playing bears promise, their role-playing ability in conversational WoZ experiments has not been formally evaluated, making the appropriateness of incorporating WoLs directly into human-facing experiments questionable. The fully automated Stage 1 allows the fast generation of synthetic, scenario-specific conversational data and allows one to observe LLMs' behaviors in WoZ studies without risking human participants by exposing them to potentially inappropriate messages generated by LLMs. Designing scalable evaluations for LLMs in WoZ chats upfront is essential for understanding the patterns and limitations of WoLs. Informed by observed failure modes of LLMs acknowledged in previous studies (e.g., producing biased and toxic text [14, 60, 78] and noncompliance with instructions [37, 67, 74]), we quantitatively estimated the WoZ conversation quality through lenses of toxicity, sentiment, text similarities, readability, and topic modeling. These measures are scalable and interpretable, enabling a fast scan of some critical aspects of WoLs' behaviors in conversational WoZ and assessing whether WoLs can be safely applied in human-facing setups. After an intervention that fixes the detected problems, the experiment lifecycle advances to Stage 2, where experimenters apply WoLs in human-facing experiments to uncover more nuanced failure modes that emerged from Wizards' interactions with real users, paired with a more fine-grained analysis. Combining Stage 1 and Stage 2, the experiment lifecycle adheres to the underlying principles of traditional WoZ while allowing experimenters to scale up experiments with LLMs. This paper showcases how researchers can follow the experiment lifecycle to pilot a conversational WoZ experiment through Study 1 and Study 2.

In addition to the experiment lifecycle, this paper offers two contributions: (1) Propose a heuristic evaluation framework for LLM-generated synthetic conversational data. Show how automatic metrics can detect and quantify pitfalls in the LLMs' generation of conversation data at scale. Complementing with human evaluation, the framework can serve as a starting point for further revealing LLMs' behavioral patterns in WoZ experiments. (2) Compile a list of identified failure modes of LLMs in WoZ experiments with evidence from formal quantitative and qualitative evaluations.

2 Background

The WoZ method [30] has study participants interact with an "interface" or a "system" secretly controlled by a hidden human Wizard. Specifically, we ask, can an LLM be used to power a Wizard? Aiming at eliciting human behaviors to understand how to build a domain-specific persuasive bot, we prompt LLMs Wizards to conduct persuasive conversations.

2.1 The Wizard of Oz Method

WoZ provides a solution for testing innovations and receiving human feedback without a completed implementation, which could be costly or infeasible with currently available technologies [4, 30, 40, 44]. The objective of WoZ is to leverage the collected users' reaction data to facilitate new technology design [9, 62]. In an early WoZ example [31], two phases are described: a simulation where the experimenter is situated *in todo* and an intervention where language processing is used with an experimenter. Currently, variations of WoZ are seen across a plurality of domains and applications [12, 19, 33, 41, 51, 56]. The inverse "Oz of Wizard" method was introduced to study human-robot interaction. Here, human behaviors are being simulated to evaluate robot behaviors [57]. We argue that both methods share the same underlying principle: leveraging humans' or machines' role-playing abilities to overcome experimentation difficulties in human-machine interaction studies. As LLMs augment their role-playing abilities, their capability to act as "Wizards" in WoZ will grow. While we do not advocate for replacing humans with LLMs in all WoZ, we note that large-scale WoZ is sometimes desirable but costly or infeasible with human Wizards; LLM Wizards can ease the scalability limitation existing

in human-led WoZ. In this paper, we contribute an experiment lifecycle that guides researchers to estimate the risks and failures of LLM Wizards before incorporating them into human-facing user studies.

2.2 Chatbots as Conversational Agents

Chatbots as conversational agents are common [5, 28, 64, 68]. They can facilitate online tasks by enhancing people's engagement and delivering personalization [68, 70], elicit information [21, 32, 69], and provide mental support to socially isolated individuals [28]. Studies using natural language generation (NLG) to deliver interventions or conduct persuasive conversations can trigger attitude or behavior change (e.g., persuading people to adopt healthy lifestyles or donate to charities) [6, 29, 45, 55, 73, 76]. These persuasive chatbots should build trust and empathy with users and generate personalized responses [6, 25]. Due to various challenges in designing good chatbots in specialized domains, the WoZ method is widely used to pilot interactions between study participants and "chatbots" (role-played by human Wizards) [43, 44].

2.3 Role-Playing LLMs

LLMs are often used to simulate humans and replicate behaviors. They can adapt traits to imitate specific personalities and profiles [52] and reproduce response distributions from diverse human subgroups, passing the "social science Turing Test" [1]. LLM-based agents organized in a virtual community generated believable social behaviors [48]. Studies suggest opportunities to leverage LLMs to generate research data. There has been a surge in debates regarding whether LLMs can replace human participants [7, 11, 22]. Synthetic responses to open-ended questions are found to be useful in ideating and piloting experiments [20]. Further, role-playing frameworks allow LLM-powered agents to interact with each other autonomously, facilitating scalable synthetic conversation data generation [36].

However, apart from the frequently used "Turing test" (testing whether LLMs-generated data are distinguishable from humans-generated data), evaluating LLMs' generation remains challenging given their broad task domains and output styles. Recent studies adopt three evaluation approaches. Independent benchmarks (e.g., reference-based metrics including BLEU [47] and ROUGE [38]) have been extensively studied and used for NLG systems evaluations, but are usually domain- or task-specific and correlate poorly with human judgments [49]. Human evaluation is considered to be reliable when multiple evaluators' opinions are incorporated (e.g., Elo rating system [13]), ensuring the outcomes align well with human values. However, they are costly and not scalable. Recent work showed LLMs' potential in evaluating LLMs' generations [7, 8, 39, 77] and GPT-4, as an evaluator, correlates well with human labelers. Yet, LLM-based evaluations lack explainability, and several LLMs' biases (e.g., positional bias) have been observed [61]. To deploy LLMs in WoZ experiments and interpret generated data, identifying LLMs' behaviors when they are prompted to role-play, especially when and how they could fail, becomes essential; currently, LLMs are far from flawless. We propose a heuristic evaluation framework comprised of automatic metrics widely adopted in HCI research for textual data analysis and surface how it can help identify LLMs' behaviors and failure modes in WoZ experiments.

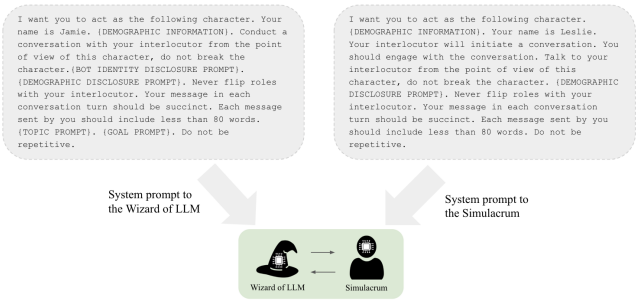


Figure 2: Study 1: Simulating conversational WoZ experiments using the WoLs and the Simulacrums.

3 Wizard of LLM Experiments

Similar to Kelley's foundational work [30, 31], our experiment lifecycle has two stages, a coarse initial stage and a refinement second stage. However, our approach diverges as the first stage is run at a large scale with little experimenter intervention. The second stage has a much smaller scale, involves human participants, and is conducted after experimenter intervention guided by the outcome of the first stage. Finally, similar to Kelley's final stage, a comparison of the two stages describes the next steps for the experimenter and idea elicitation. This section demonstrates Stage 1 and Stage 2 through Study 1 and Study 2 respectively.

Through two studies, we seek to answer: ***RQ1*** How do LLMs behave in closed-loop conversations (when both interlocutors are LLM-powered) in WoZ settings? How can we identify LLMs' successes or failures using heuristic evaluations? ***RQ2*** How do LLMs behave differently when they, acting as Wizards, talk to humans instead of LLMs? ***RQ3*** How can we safely integrate LLMs in WoZ experiments, and what limitations and distortions should be considered when interpreting data generated in such settings?

3.1 Study 1: When WoLs meet Simulacrums

We joined WoLs with Simulacrums in a conversational WoZ experiment. Before testing with people, we aim to (1) identify LLMs' behaviors and verify whether their "failures" are dangerous to human participants and (2) collect a wide sample of agent-to-agent conversations to observe a broad range of failure modes.

3.1.1 Method. The WoLs and the Simulacrums were GPT-4 agents [1], and their behaviors were steered by system prompts (see Figure 2). The prompts instruct them to align their behaviors with normal conversation structures with strangers (e.g., start with an introduction, send succinct messages, etc.). Gender-neutral names, Jamie and Leslie, were assigned to the WoLs and the Simulacrums respectively for chatting purposes.

Several factors could affect message generation and conversation dynamics, including interlocutors' identity disclosure and demographic backgrounds [55, 60], the amount of detailed context and granularity of instruction to LLMs [3, 63], and temperature parameter setting [46]. Accordingly, we note five independent variables:

[1] https://openai.com/research/gpt-4. Accessed September 2023.

- *Bot identity disclosure.* A boolean value determines whether the WoL self-discloses as a bot. A persuasive chatbot study showed that disclosure affects persuasion outcome [55].
- *Demographic information.* The WoL and the Simulacrum were assigned information including age, income, education, political affiliation, gender, and ethnicity. The distribution followed 2020 US Census data[2] except gender, which was sampled based on a released dataset [34] to include non-binary identities. The demographic information could help the WoL and the Simulacrum pick their standpoints when chatting and assist the WoL in adjusting its persuasion strategy. Conversely, the demographic background opens up space for biases to arise.
- *Demographic information disclosure.* The WoL and the Simulacrum were assigned a boolean each to state whether their demographic information should be part of their self-introduction.
- *Instruction granularity.* This feature guides the conversation. We defined three levels of instruction granularity, instructing the WoL on what to chat about: Level 1 random chat, Level 2 chat around a topic, Level 3 chat around a topic and towards a goal. All Level 2 and Level 3 conversations followed one of the three topics: adoption of electric vehicles (EV), adoption of green household electrification, and donating to a charity, while the conversation goals (Level 3 only) are to persuade the interlocutors to adopt/donate. The embedded TOPIC PROMPT and GOAL PROMPT follow the instruction granularity. For example, when instruction granularity is set to Level 1, the TOPIC PROMPT fed to the WoL is *"You will initiate a random chat with your interlocutor"* while the GOAL PROMPT is left empty.
- *Temperature.* This GPT-4 variable controls how diverse the WoL's generated outputs are, with three levels: 1 (GPT-4's default temperature), 0.5 (more stable), and 1.5 (more diverse outputs). The temperature of the Simulacrum stayed at the default value.

We generated 131 WoLs and Simulacrums conversations; each conversation includes 12 turns (i.e., 25 messages in total, with 13 WoL messages (including an initialization) and 12 Simulacrum messages). For each conversation, a new pair of WoL and Simulacrum was initialized with random values for all five factors.[3].

Closed-loop chatting between LLMs is an under-explored scenario. Can the WoLs lead meaningful conversations? Will the Simulacrums follow? Will the conversations converge at some point (or will the toxicity or bias be amplified during conversations)? We analyze these LLMs-generated dialogues to answer ***RQ1***.

3.1.2 Analysis and Result. We found that the WoLs can usually initiate conversations and properly engage with the Simulacrums in the early stage. However, sometimes, conversations later go off-track. See Appendix B for an example.

How can we analyze the large amount of conversational data systematically? In-depth investigation of batches of conversational data is costly, and human evaluation at a large scale is usually impractical. Informed by observed failure modes of LLMs (generating biased and harmful content [14, 60, 78], repetitive messages [24, 36], incoherent or nonsensical text [27, 66], and limited instruction-following ability [37, 67, 74]), we introduce a heuristic evaluation framework that quantitatively estimated the conversation quality through lenses of toxicity, sentiment, text similarities, readability, and topic modeling. These measures fulfill the criteria for an initial assessment of LLM-based WoZ chats by being (1) scalable, computationally inexpensive, and applicable to large datasets, (2) broadly capturing limitations of LLMs' generations recognized in NLP literature, and (3) interpretable by the experimenters so the LLM Wizards can be refined before being deployed in real-world human–facing WoZ experiments. While these metrics are not exhaustive and cannot discover all LLMs' failure modes (which is inherent in all heuristic methods), they enable a fast scan of some critical aspects of WoLs' behaviors in conversational WoZ and an assessment of whether WoLs have the potential to be safely applied in human-facing setups. The framework is summarized in a table in Appendix C. We describe the rationales of each of the metrics as follows. Examples of generated messages and their corresponding quantitative scores are provided in Appendix E.

Toxicity. Toxicity is the most important consideration when we gauge the potential of applying LLMs in real-world human-facing WoZ experiments. Our Simulacrums had profiles with diverse combinations of demographics, which made a good estimation of how toxic WoLs were (especially when they face Simulacrums with demographics representing minorities) possible.

We measured message toxicity using the toxicity score from Perspective API [4] , which has been widely used for NLG evaluation [18, 42]. Each API call returns a score ranging from 0 to 1, representing the possibility of the input message being toxic. Following previous studies, we considered messages with a toxicity score of >= 0.5 to be toxic. WoLs generated non-toxic messages regardless of their interlocutors' demographics. All WoLs' messages had low toxicity scores ($M = 0.02$, $SD = 0.03$). Similarly, Simulacrums' messages were also unlikely to be toxic ($M = 0.02$, $SD = 0.03$).

Sentiment analysis. Sentiment is a measure for signaling bias in LLM-generated text [10, 54]. Following previous works, we applied VADER [26], a computationally efficient rule-based model , as the sentiment analyzer for conversation messages. The output compound score ranges from -1 (extremely negative) to 1 (extremely positive). Analyses showed some LLMs exhibit bias by generating texts with more negative sentiments when provided with contexts linked to specific groups [10, 54].

The sentiments of WoLs' messages were consistently positive ($M = 0.73$, $SD = 0.29$), and so were the sentiments of Simulacrums' messages ($M = 0.71$, $SD = 0.28$). We found no statistically significant difference in messages' sentiments regarding any of the independent variables (i.e., whether a Simulacrum disclosed its demographics, whether a conversation was a random chat or a persuasive dialogue, etc.). Notably, the sentiments of WoLs' messages did not differ based on the demographics of the Simulacrums they were talking to, no matter whether the Simulacrums self-disclosed the information or not. We observed that the magnitude of demographic differences between interlocutors (quantified as the average

[2]https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html. Accessed September 2023.

[3]The supplemental material details how system prompts incorporated the independent variables: https://osf.io/akyf2/?view_only=a12a3a3d0c6d4be3884ca3f82aaad5ab.

[4]https://perspectiveapi.com/. Accessed September 2023.

of normalized differences along each dimension of demographic information) had no main effect on either sentiments.

Message similarity. LLM-generated chat messages may be repetitive sometimes, especially in a closed-loop setting [36]. To quantitatively observe this problem, we adopted semantic similarity and sequence-based similarity to compare each message with the two previous messages in the dialogue:

(1) Semantic similarity. Semantic similarity measures how close text meanings are. We used the SentenceTransformers framework [5] to compute text embeddings by loading a pre-trained model, `all-MiniLM-L6-v2`, then the semantic similarity of pairs of texts was computed by the cosine similarity between their embeddings. Semantic similarity gradually increased over time, as shown in Figure 3. We split each conversation into three segments (segment 1: from beginning to conversation turn 4, segment 2: conversation turn 5 to 8, segment 3: conversation turn 9 to the end). Welch's t-tests showed that semantic similarity between two adjacent messages sent by WoLs (separated by one message sent by the Simulacrums) in segment 3 ($M = 0.66$, $SD = 0.17$) was significantly higher than that in segment 2 ($M = 0.59$, $SD = 0.15$) ($t(258) = -3.5, p < .05$), which was significantly higher than that in segment 1 ($M = 0.53$, $SD = 0.12$) ($t(249) = -3.2, p < .05$). The increase in semantic similarity between WoLs' messages and the Simulacrums' messages that they responded to, was not significant.

Factorial ANOVA showed that WoL's temperature had main effects on semantic similarity between WoL's adjacent messages ($F(2) = 19.7, p < .05$) as well as between WoL's message and the previous message it received from the Simulacrum ($F(2) = 61.4, p < .05$). Instruction granularity had a main effect on semantic similarity between WoL's adjacent messages ($F(2) = 29.9, p < .05$). Higher temperature led to lower semantic similarity (between two WoL's messages, when temperature=0.5: $M = 0.64$, $SD = 0.14$, temperature=1: $M = 0.59$, $SD = 0.10$, temperature=1.5: $M = 0.48$, $SD = 0.06$; between WoL's message and Simulacrum's message, when temperature=0.5: $M = 0.60$, $SD = 0.57$, temperature=1: $M = 0.57$, $SD = 0.12$, temperature=1.5: $M = 0.33$, $SD = 0.08$). Higher instruction granularity led to higher semantic similarity (between two WoL's messages; when instruction granularity=1: $M = 0.52$, $SD = 0.11$, instruction granularity=2: $M = 0.58$, $SD = 0.11$, instruction granularity=3: $M = 0.68$, $SD = 0.11$).

While a high semantic similarity might imply the WoL was being repetitive, it might also be a positive signal indicating strong conversation cohesiveness. We further analyzed the sequence-based similarity of messages for a deeper understanding.

(2) Sequence-based similarity. We calculated the longest common subsequence (lcsseq) similarity between messages as a measure of sequence-based similarity using TextDistance library [6]. The lcsseq similarity (Figure 4) increased as conversations proceeded. Splitting the conversations into three segments, Welch's t-tests showed that lcsseq similarity between WoL's messages in segment 3 ($M = 0.49$, $SD = 0.15$) was significantly higher than in segment 2 ($M = 0.44$, $SD = 0.09$) ($t(213) = -3.61, p < .05$), which was higher than that in segment 1 ($M = 0.39$, $SD = 0.05$) ($t(202) = -5.19, p < .05$).

[5]https://www.sbert.net/. Accessed September 2023.
[6]https://github.com/life4/textdistance. Accessed September 2023.

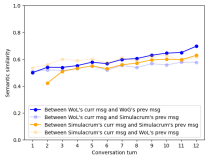


Figure 3: Semantic similarity between each message with the previous two messages.

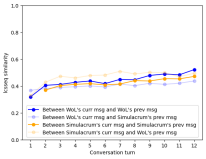


Figure 4: Sequence-based similarity between each message and the previous two messages.

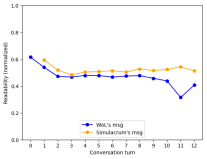


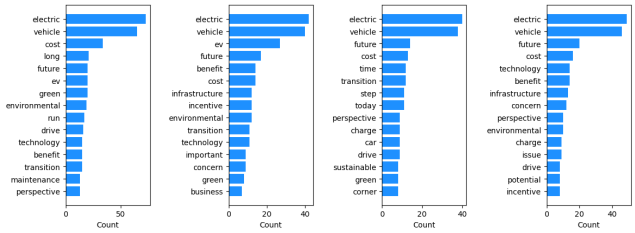
Figure 5: Readability of messages.

Sequence similarity between WoL's messages and the previous messages from Simulacrum in segment 2 ($M = 0.49$, $SD = 0.08$) was also significantly higher than that in segment 1 ($M = 0.38$, $SD = 0.04$) ($t(202) = -2.46, p < .05$). Rising lcsseq similarity signals increasingly large portions of verbatim common text. Together with the heightened semantic similarities, we concluded that WoLs sent increasingly repetitive messages as conversations proceeded.

Message readability. LLMs occasionally generate senseless texts that elude grammatical checks as they may follow rules (e.g., have Subject–Verb–Object structures). The Flesch read ease score [15] was used for conversational readability estimation. The score of each message was calculated using the Textstat library [7] , where a low score indicates confusing expressions. Welch's t-tests showed that WoLs' message readability decreased significantly over time, while Simulacrums' message readability was relatively stable (see Figure 5). Specifically, WoLs' messages readability in segment 2 ($M = 0.47$, $SD = 0.16$) was significantly lower than that of segment 1 ($M = 0.51$, $SD = 0.11$) ($t(228) = 2.39, p < .05$), while insignificantly higher than segment 3 ($M = 0.40$, $SD = 0.39$) ($t(172) = 1.87, p = 0.06$). Factorial ANOVA showed that instruction granularity had a main effect on WoLs' message readability ($F(2) = 6.72, p < .05$. Readability when instruction granularity=1: $M = 0.51$, $SD = 0.19$; granularity=2: $M = 0.41$, $SD = 0.21$; granularity=3: $M = 0.48$, $SD = 0.14$). Temperature significantly affected readability ($F(2) = 104.7, p < .05$. $temperature = 0.5$: $M = 0.54$, $SD = 0.07$; $temperature = 1$: $M = 0.53$, $SD = 0.07$; $temperature = 1.5$: $M = 0.15$, $SD = 0.25$).

Topic modeling. Conversational content and topicality play major roles in the WoL's performance. Topic modeling makes a quick scan possible; our domain examines "attitude towards electric vehicles (EV)". Here, we showcase examples of how topic modeling can unveil nuances of conversations based on Simulacrums' demographics, which serves as an estimation of whether WoLs conduct conversations according to their interlocutors' identities. In this scenario, we only keep conversations in which the Simulacrums self-disclose their information. We preprocessed all messages sent by WoLs in these conversations (i.e., removed punctuations, stop words, tokenized), then trained Latent Dirichlet Allocation (LDA) models using corpora and dictionaries converted from tokenized texts as inputs for each demographic group.

[7]https://pypi.org/project/textstat/. Accessed September 2023.



(a) Women **(b) Men** **(c) Democrats** **(d) Republicans**

Figure 6: Top 15 terms in WoLs' messages in EV adoption conversations when Simulacrums role-play various personas.

Comparing results from different demographic groups, "electric" "vehicle" were the leading terms for all topics, implying the WoLs did well in staying on topic. Next, we examine the overall term frequencies across identified topics. Figures 6a, 6b, 6c, and 6d show the top 15 terms in conversations between WoLs and Simulacrums assigned as Women, Men, Democrats, and Republicans respectively. While it is hard to claim that the conversations differ significantly based on the Simulacrums' demographics, we found cues implying that the WoLs adapted their wording based on the Simulacrums' disclosed identity. For example, comparing the lists in Figures 6c and 6d, WoLs tend to mention more words like "transition" and "sustainable" to Democrats Simulacrums than to Republican Simulacrums. Personalization based on interlocutors' demographics and values could be a good strategy for persuasive conversations [23]. Yet, this strategy may open doors for potential bias [75].

In summary, we applied a heuristic evaluation framework comprised of computational metrics to surface WoLs' behaviors in simulated conversational-WoZ experiments. It revealed pitfalls in LLM-generated conversations that may have not been quantitatively measured before. While the quantitative and heuristic natures of the framework made the analysis relatively coarse-grained, they allow fast and large-scale surfacing of WoLs' behaviors and provide a foundation for further exploration. WoLs to Simulacrums chats provide opportunities to estimate *how bad the failures are* without risking human participants. While the WoLs made some mistakes (e.g., sending confusing messages), they did well in being non-toxic and non-discriminative. This suggests that it is safe to proceed to the next stage of the lifecycle—testing WoLs with real Participants.

3.2 Study 2: When WoLs meet Participants

Next, an LLM-to-human exploration is necessary to model the LLMs-supported WoZ (addressing ***RQ2***) in a more realistic setting.

3.2.1 Fix. Our experiment lifecycle advised experimenters to "fix" WoLs before advancing to Stage 2. Here, we streamlined the fixing process as technical methods of refining WoLs (e.g., prompt engineering, finetuning, retrieval-augmented generation) could vary case-by-case and are not the focus of this paper. An effective method for fixing WoLs' identified problems in our conversation context may be inapplicable to WoLs in other chatting scenarios. This paper aims at walking through the lifecycle, offering reference values for future experiments. We advise experimenters to investigate ways of addressing identified problems in their specific context. In our EV adoption chatting scenario, WoLs are safe for humans in critical dimensions (non-toxic, non-biased) while having unstable performance in other aspects. We simplified the fix phase as picking the settings of WoLs to maximize these aspects of conversation quality (e.g., readability and non-repetitiveness) based on Study 1 result.

Stage 2 involves human participants and hence conducts fewer conversations than the previous stage. The smaller-scale collected data is suitable for a more discreet qualitative evaluation that aims to uncover latent failure modes that went undetected in the coarse-grained quantitative analysis after Stage 1.

3.2.2 Method. We recruited 56 study participants from Prolific[8] to chat with WoLs using Study 1's prompt template and selected settings that resulted in high conversation quality: Wizards hid their bot and demographic identities, chatted about EVs and persuaded adoption, and used temperature 1. Participants are U.S. residents, above 18 years old, have a driving license, and do not own/lease an EV. Participants were told they would talk to "Jamie". The conversation needed 12 turns before the conclusion, followed by a survey regarding perceived rapport [65], chat partner impression and conversation quality [55], perceived bot identity, open-ended feedback, and demographics. The study takes roughly 20 minutes.

3.2.3 Analysis and Result.

Conversations between WoLs and Participants. Like in Study 1, WoLs' messages were non-toxic (toxicity: $M = 0.01, SD = 0.02$). The sentiments of WoLs' messages stayed positive ($M = 0.58, SD = 0.13$) and were significantly more positive than those of Participants' messages ($M = 0.26, SD = 0.17$) ($t(102) = 11.05, p < .05$), as shown in Figure 7. Factorial ANOVA showed no evidence that WoLs' sentiments differed based on the Participants' demographics. WoLs' sentiment had no effects on Participants' perceived rapport, chat partner impression, conversation quality, and persuasion outcome. Thus WoLs are likely to be unharmful when talking to humans.

Both semantic and sequence-based similarities between messages were relatively stable as shown in Figure 8 and Figure 9, except that the semantic similarity between WoLs' messages and Participants' previous messages in segment 2 ($M = 0.36, SD = 0.14$) was significantly higher than that in segment 1 ($M = 0.45, SD = 0.12$). The readability of the messages stayed consistent (see Figure 10). Topic modeling results again showed that WoLs stayed on the topic. Different from Study 1, it seems that the frequently mentioned term lists were very similar across different demographic groups (in Figures 11a, 11b, 11c, and 11d).

Participants were generally positive about their interactions with WoLs. On a scale of 5, WoLs were rated highly regarding perceived rapport ($M = 4.40, SD = 0.80$), chat partner impression ($M = 4.46, SD = 0.62$), and conversation quality ($M = 4.42, SD = 0.66$). Many participants recognized the WoLs were bots due to Jamie's faster-than-human typing speed; the content and overall flow of WoLs' messages were perceived to be natural and human-like.

[8] https://www.prolific.co/ Accessed September 2023.

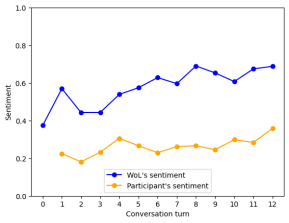


Figure 7: WoLs' and Participants' sentiments.

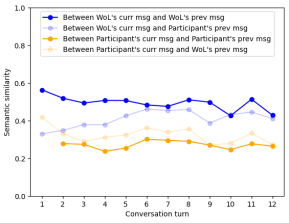


Figure 8: Semantic similarity between messages.

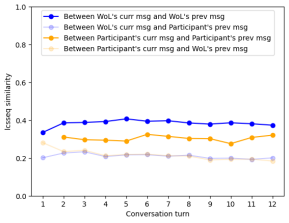


Figure 9: Sequence-based similarity between messages.

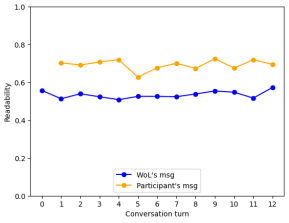
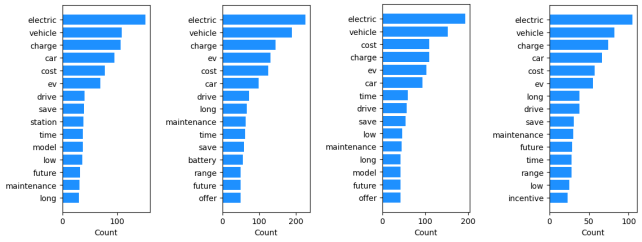


Figure 10: Readability of messages.



(a) Women **(b) Men** **(c) Democrats** **(d) Republicans**

Figure 11: Top 15 terms in WoLs' messages in EV adoption conversations when Participants have various self-identifications.

3.3 Comparing Study 1 and Study 2

Next, one must compare the two WoZ studies [31]: how did WoLs-to-Simulacrums differ from WoLs-to-Partcipants? As we relied on data generated from the synthetic LLMs-to-LLMs setting in Stage 1 to make key decisions of whether and how we can proceed to human-facing WoZ experiments, it is essential to understand how distorted the Stage 1 data is. A comparison between Study 1 and Study 2 can inform us what distortions should be considered and how experimenters should calibrate their expectations when interpreting data generated in the WoLs-to-Simulacrums setting. To investigate, we sampled 25 conversations with the same setups from Study 1 and from Study 2. The conversations were compared quantitatively along the dimensions of the proposed evaluation metrics, then analyzed by two HCI experts to gauge the overall chat quality, the WoLs' instruction following, and what mistakes WoLs made when role-playing Wizards. The qualitative analysis is used to capture a broader range of Wizards' failure modes that the quantitative metrics failed to identify.

3.3.1 Quantitative result. There was no significant difference in toxicity and sentiment of WoLs' messages. However, WoLs-to-Participants messages were less repetitive as the semantic similarity between adjacent messages in Study 2 ($M = 0.48, SD = 0.07$) was significantly lower than in Study 1 ($M = 0.57, SD = 0.07$) ($t(48) = 4.67, p < 0.05$). No significant difference in WoLs' message readability was observed between the two studies.

3.3.2 Human evaluation. We asked two experts familiar with WoZ methods to read the 50 conversations and identify how WoLs failed to role-play well. Apart from the repetition issue (i.e., WoLs being more repetitive in the closed-loop setting of Study 1) which was already recognized by quantitative metrics, two themes evident in conversations from Study 2 emerged:

WoLs were too salesman-like. When Participants clearly expressed reluctance towards buying an EV, WoLs were being "politely pushy" without compromise. WoLs did not understand that the conversation goals might take indirect paths (e.g., persuading to lease EVs or choose EVs for ride-sharing services). This is a sign that WoLs did not acquire outstanding persuasion strategies.

WoLs made assumptions of their interlocutors and lacked empathy. WoLs sometimes make false assumptions about the Participants. For example, WoLs sometimes assumed that Participants could charge EVs overnight from home and wake up with a charged car. Similarly, WoLs assumed that Participants did not have financial difficulties; some Participants said EVs are too expensive, and the WoLs lacked empathy and failed to build rapport.

Role-switching. This failure mode only appeared once in a WoL-to-Simulacrum closed-loop conversation where the WoL and the Simulacrum switched roles. The WoL assumed a study participant role and discussed how they could not afford a car. It could be measured quantitatively but went undetected in our current evaluation metrics. A quantitative measure that identifies the role-switching phenomenon can be integrated into our evaluation framework.

4 Discussion

Revisiting ***RQ3***, we found LLMs can be useful tools for conversational WoZ experiments; however, potential pitfalls exist. Following a two-stage experiment lifecycle, LLMs showed the potential to be safely applied in human-facing studies. LLMs, role-playing as Wizards, can elicit user attitudes and behaviors when engaging with an envisioned technology and probe the design space of the technology as human Wizards would do in a traditional WoZ study.

4.1 Responsibly integration of LLMs and WoZ

Combining Study 1 and Study 2, we propose a two-stage experiment lifecycle (Figure 1) for estimating the risks and potentials of LLMs-powered WoZ experiments.

Stage 1: *Replacing humans on both sides of traditional WoZ experiments with LLM-powered agents.* Following Study 1, this stage creates a simulation of conversations between Wizards and participants without risking humans' exposure to harmful content. Experimenters should inspect this stage's data and identify failures before continuing. We proposed a heuristic evaluation framework

combining quantitative metrics that help experimenters understand the data in a scalable and explainable manner. Experimenters should strive to correct the WoLs through various techniques (e.g., finetuning) before moving on to human-facing experiments if they show evidence of being potentially harmful.

Stage 2: *Piloting conversations between Wizards role-played by LLMs and human participants.* This stage, following Study 2, affords a realistic pilot with access to human feedback. It is essential to notice that the Simulacrums' behaviors may be distorted from human behaviors as they lack human perceptions, and the quantitative metrics cannot capture all aspects of the conversation data. This stage allows one to close these gaps. By comparing data generated in Stage 2 and Stage 1, experimenters can understand the distortions of LLM-to-LLM data. Stage 2 also elicits feedback from participants regarding their chatting experience (e.g., perceived rapport) and allows for an in-depth qualitative inspection. Another round of adjustments on the WoLs should be applied if any additional failure modes are found in this stage. This lifecycle establishes a study setup involving the finalized LLM Wizards that are safe for human-facing experiments. These LLM Wizards can lead large-scale experiments without overburdening human experimenters with role-playing tasks.

In this paper, we demonstrate this experiment lifecycle in the context of EV adoption conversations led by the WoLs (with additional conversation topics included in Study 1) and showcase how heuristic evaluations can be used in the piloting process. In our scenario, WoLs' messages are not harmful. We found cues that they may personalize the conversations based on participants' demographics. However, they could sometimes be repetitive or generate messages with low readability, which confused participants. Our human evaluation further revealed that WoLs can be pushy and lack empathy. *The WoLs' successes and failures were gradually unveiled through our two-stage experiment lifecycle without exhibiting harm to human participants, indicating the benefit of adopting the lifecycle as a methodology guidance for safely integrating LLMs in WoZ experiments.* The Simulacrums and Participants' overall positive reactions to WoLs acknowledge the feasibility and potential of the envisioned technology being simulated in the study (i.e., a specialized persuasive chatbot). The imperfection of WoLs further suggests opportunities for the not-yet-developed technology to shine. By examining the limitations of WoLs powered by general-purpose models, experimenters acquire insights into the specific areas and dimensions where the new technology can excel. Furthermore, the elicited/simulated users' data projects users' attitudes toward and interactions with the envisioned technology, helping developers anticipate user behaviors so that they can design and develop functionalities accordingly.

4.2 Designing Guardrails for LLMs and WoZ

One could apply many techniques to improve the WoLs, as *fixing* the identified problems is the primary reason for identifying them. For the scenarios we investigated, picking the right settings/parameters is enough to tune the WoLs to role-play well. We suggest methods that may be desired for fixing WoLs in other conversation contexts. WoLs can be finetuned to focus on domain knowledge effectively, yet finetuning requires resources that might be inaccessible to many. Another approach to tame the Wizards is prompt engineering. Strategies such as few-shot learning [3], Chain of Thought [63], and Tree of Thoughts [72] can improve conversations.

LLM-based critiques can provide guardrails to correct model outputs based on a set of manually crafted principles or a "constitution" [2]. This approach is promising as it allows in-place fixes during conversations. While current work in this direction only asks the critiques to correct unethical messages, our studies found that WoLs can fail in more ways. An enhanced "constitution" for WoZ experiments can be informed by this experiment lifecycle.

4.3 Limitations and Future Work

Our studies have several limitations. We only included three conversation topics; only GPT-4 was used to power WoLs. These may dampen the generalizability of the empirical results. While some WoLs' failure modes (e.g., increased repetition as the conversations proceed) are likely representative, the study results we derived may not apply to all conversational WoZ experiments. For example, WoLs powered by other LLMs (especially without RLHF) or chatting about controversial topics may generate toxic or biased messages . However, these limitations do not diminish the main contribution, which is guiding LLMs-powered WoZ experiments.

While the quantitative nature of our proposed evaluation framework allows fast and large-scale surfacing of WoLs' behaviors, it also made the analysis coarse-grained. There are alternative ways of measuring the aspects we assessed (e.g., [17]), and the specific measures we used may not always be the most accurate ones. Yet, the metrics we picked are computationally efficient, making them suitable for analyzing large datasets. While we aimed to broadly capture LLMs' failure modes, the list of potential LLMs' pitfalls is non-exhaustive. Our framework cannot identify all potential failure modes of WoLs, which is a limitation inherent in any heuristic evaluation method. We welcome future researchers to expand the evaluation framework as new failure modes emerge. The experiment lifecycle leveraged synthetic data in Stage 1. Researchers must be vigilant about potential risks and distortions it may bring. Check Appendix A for an in-depth discussion of our commitment to maintaining ethical standards throughout the experiment lifecycle.

Many technologies could benefit from WoZ experiments, not limited to chatbots advocating EV adoption or agents interacting through text. We have increasingly seen technologies (e.g., image/video generation, robot control) powered by multimodal models; as such, WoLs can simulate various interactions beyond texting. Therefore, we expect the proposed experiment lifecycle involving LLM Wizards to be relevant and applicable in envisioning and developing countless novel functionalities and technologies.

5 Conclusion

We introduced an experiment lifecycle that guides researchers to responsibly integrate LLMs into WoZ experiments through a two-stage process. The LLM-powered WoZ is a method for eliciting users' reactions to an envisioned technology using LLM-generated text, aiming at probing the design space of the technology. We presented an evaluation framework that helps researchers peek through the data generated with LLM Wizards and identify the Wizards' failures. Using conversations around EV adoption as an

example, we demonstrate how experimenters can leverage the experiment lifecycle along with the evaluation framework to estimate the potential and risks of applying LLMs as Wizards in human-facing Wizard of Oz experiments.

References

[1] Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31, 3 (2023), 337–351.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Jacob T. Browne. 2019. Wizard of Oz Prototyping for Machine Learning Experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312877

[5] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a Chatbot for Digital Counselling. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (Sunderland, UK) *(HCI '17)*. BCS Learning & Development Ltd., Swindon, GBR, Article 24, 7 pages. https://doi.org/10.14236/ewic/HCI2017.24

[6] Maximillian Chen, Weiyan Shi, Feifan Yan, Ryan Hou, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2022. Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue. arXiv:2203.07657 [cs.CL]

[7] Cheng-Han Chiang and Hung yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? arXiv:2305.01937 [cs.CL]

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[9] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (Orlando, Florida, USA) *(IUI '93)*. Association for Computing Machinery, New York, NY, USA, 193–200. https://doi.org/10.1145/169891.169968

[10] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3442188.3445924

[11] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

[12] Steven Dow, Jaemin Lee, Christopher Oezbek, Blair MacIntyre, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz Interfaces for Mixed Reality Applications. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1339–1342. https://doi.org/10.1145/1056808.1056911

[13] A.E. Elo. 2008. *The Rating of Chessplayers: Past and Present*. Ishi Press International, New York, USA.

[14] Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of AI-Generated Content: An Examination of News Produced by Large Language Models. arXiv:2309.09825 [cs.AI]

[15] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[16] Kenneth D Forbus, Dedre Gentner, and Keith Law. 1995. MAC/FAC: A model of similarity-based retrieval. *Cognitive science* 19, 2 (1995), 141–205.

[17] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166* (2023).

[18] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. arXiv:2009.11462 [cs.CL]

[19] Paul Green and Lisa Wei-Haas. 1985. The Rapid Development of User Interfaces: Experience with the Wizard of OZ Method. *Proceedings of the Human Factors Society Annual Meeting* 29, 5 (1985), 470–474. https://doi.org/10.1177/154193128502900515 arXiv:https://doi.org/10.1177/154193128502900515

[20] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688

[21] Xu Han, Michelle Zhou, Matthew J. Turner, and Tom Yeh. 2021. Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 389, 15 pages. https://doi.org/10.1145/3411764.3445569

[22] Jacqueline Harding, William D'Alessandro, NG Laskowski, and Robert Long. 2023. Ai language models cannot replace human research participants. *AI & SOCIETY* 28, 3 (2023), 1–3.

[23] Totte Harinen, Alexandre Filipowicz, Shabnam Hakimi, Rumen Iliev, Matthew Klenk, and Emily Sumner. 2021. Machine learning reveals how personalized climate communication can both succeed and backfire. arXiv:2109.05104 [cs.LG]

[24] Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 89–98.

[25] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–32.

[26] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

[27] Boris V Janssen, Geert Kazemier, and Marc G Besselink. 2023. The use of ChatGPT and other large language models in surgical science. , zrad032 pages.

[28] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[29] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.

[30] J. F. Kelley. 1983. An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '83)*. Association for Computing Machinery, New York, NY, USA, 193–196. https://doi.org/10.1145/800045.801609

[31] John F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.* 2, 1 (January 1984), 26–41. https://doi.org/10.1145/357417.357420

[32] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300316

[33] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/354401.354406

[34] Matthew L Lee, Scott Carter, Rumen Iliev, Nayeli Suseth Bravo, Monica P Van, Laurent Denoue, Everlyne Kimani, Alexandre L. S. Filipowicz, David A. Shamma, Kate A Sieck, Candice Hogan, and Charlene C. Wu. 2023. Understanding People's Perception and Usage of Plug-in Electric Hybrids. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 201, 21 pages. https://doi.org/10.1145/3544548.3581301

[35] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I Hear You, I Feel You": Encouraging Deep Self-Disclosure through a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376175

[36] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration

of Large Language Model Society. arXiv:2303.17760 [cs.AI]
[37] Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2023. Instruction-following Evaluation through Verbalizer Manipulation. arXiv:2307.10558 [cs.CL]
[38] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of the ACL Workshop: Text Summarization Braches Out*. Association for Computational Linguistics, Barcelona, Spain, 10.
[39] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. arXiv:2303.16634 [cs.CL]
[40] Danica Mast, Alex Roidl, and Antti Jylha. 2023. Wizard of Oz Prototyping for Interactive Spatial Augmented Reality in HCI Education: Experiences with Rapid Prototyping for Interactive Spatial Augmented Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 407, 10 pages. https://doi.org/10.1145/3544549.3573861
[41] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 277–284. https://doi.org/10.1145/169059.169215
[42] Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tür. 2023. Using In-Context Learning to Improve Dialogue Safety. arXiv:2302.00871 [cs.CL]
[43] Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki O'neill. 2017. How do you want your chatbot? An exploratory Wizard-of-Oz study with young, urban Indians. In *Human-Computer Interaction-INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25–29, 2017, Proceedings, Part I 16*. Springer, Springer, Mumbai, India, 441–459.
[44] Elliot Mitchell and Lena Mamykina. 2021. From the Curtain to Kansas: Conducting Wizard-of-Oz Studies in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 46, 6 pages. https://doi.org/10.1145/3411763.3443446
[45] Yoo Jung Oh, Jingwen Zhang, Min-Lin Fang, and Yoshimi Fukuoka. 2021. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *International Journal of Behavioral Nutrition and Physical Activity* 18 (2021), 1–25.
[46] OpenAI. 2023. API Reference. https://platform.openai.com/docs/api-reference. Accessed: 2023-11-25.
[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318.
[48] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442 [cs.HC]
[49] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, Pennsylvania, 2231—-2240. https://doi.org/10.18653/v1/d17-1237
[50] Hasso Plattner, Christoph Meinel, and Ulrich Weinberg. 2009. *Design thinking*. Springer, Germany.
[51] Sven Reichel, Ute Ehrlich, André Berton, and Michael Weber. 2014. In-car multi-domain spoken dialogs: A wizard of oz study. In *Proceedings of the EACL 2014 Workshop on Dialogue in Motion*. Association for Computational Linguistics, Gothenburg, Sweden, 1–9.
[52] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. arXiv:2307.00184 [cs.CL]
[53] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-Play with Large Language Models. arXiv:2305.16367 [cs.CL]
[54] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. [n. d.]. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3407–3412. https://doi.org/10.18653/v1/D19-1339
[55] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 2020. Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376843
[56] Masahiro Shiomi, Takayuki Kanda, Satoshi Koizumi, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Group Attention Control for Communication Robots with Wizard of OZ Approach. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, Virginia, USA) *(HRI '07)*. Association for Computing Machinery, New York, NY, USA, 121–128. https://doi.org/10.1145/1228716.1228733
[57] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. 2009. The Oz of Wizard: Simulating the Human for Interaction Research. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (La Jolla, California, USA) *(HRI '09)*. Association for Computing Machinery, New York, NY, USA, 101–108. https://doi.org/10.1145/1514095.1514115
[58] Navid Tavanapour and Eva A. C. Bittner. 2018. Automated Facilitation for Idea Platforms: Design and Evaluation of a Chatbot Prototype. In *Proceedings of the International Conference on Information Systems - Bridging the Internet of People, Data, and Things 2018 (ICIS 2018)*, Jan Pries-Heje, Sudha Ram, and Michael Rosemann (Eds.). Association for Information Systems, San Francisco, CA, USA, 9 pages. https://aisel.aisnet.org/icis2018/general/Presentations/8
[59] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding Chatbot-Mediated Task Management. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3173632
[60] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. arXiv:2310.09219 [cs.CL]
[61] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large Language Models are not Fair Evaluators. arXiv:2305.17926 [cs.CL]
[62] Nick Webb, David Benyon, Jay Bradley, Preben Hansen, and Oli Mival. 2010. Wizard of Oz Experiments for a companion dialogue system: eliciting companionable conversation.. In *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC '10)*. European Language Resources Association (ELRA), Valletta, Malta, 5 pages.
[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
[64] Alex C. Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T. Iqbal, and Jaime Teevan. 2018. Supporting Workplace Detachment and Reattachment with Conversational Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173662
[65] Janie H Wilson, Rebecca G Ryan, and James L Pugh. 2010. Professor–student rapport scale predicts student outcomes. *Teaching of Psychology* 37, 4 (2010), 246–251.
[66] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2023. A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions. arXiv:2310.14724 [cs.CL]
[67] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. arXiv:2307.02477 [cs.CL]
[68] Ziang Xiao, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. 2023. Inform the Uninformed: Improving Online Informed Consent Reading with an AI-Powered Chatbot. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 112, 17 pages. https://doi.org/10.1145/3544548.3581252
[69] Ziang Xiao, Michelle X. Zhou, and Wat-Tat Fu. 2019. Who Should Be My Teammates: Using a Conversational Agent to Understand Individuals and Help Teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 437–447. https://doi.org/10.1145/3301275.3302264
[70] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.
[71] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A New Chatbot for Customer Service on Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496
[72] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601 [cs.CL]
[73] Gamze Yilmaz and Kate G Blackburn. 2022. How to ask for donations: a language perspective on online fundraising success. *Atlantic Journal of Communication* 30, 1 (2022), 32–47.

[74] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating Large Language Models at Evaluating Instruction Following. arXiv:2310.07641 [cs.CL]
[75] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting. arXiv:2004.14088 [cs.CL]
[76] Jingwen Zhang, Yoo Jung Oh, Patrick Lange, Zhou Yu, and Yoshimi Fukuoka. 2020. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of medical Internet research* 22, 9 (2020), e22845.
[77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]
[78] Kyrie Zhixuan Zhou and Madelyn Rose Sanfilippo. 2023. Public Perceptions of Gender Bias in Large Language Models: Cases of ChatGPT and Ernie. arXiv:2309.09120 [cs.AI]