NewsInterview: a Dataset and a Playground to Evaluate LLMs' Ground Gap via Informational Interviews

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in generating coherent text but often struggle with grounded language and strategic dialogue. To address this gap, we introduce a realistic game-playing environment to facilitate the iterative improvement of dialogue agents. We focus on journalistic interviews, a domain rich in grounding communication and abundant in data. We curate a dataset of 40,000 two-person informational interviews from NPR and CNN, which we will publicly release, and reveal that LLMs are significantly less likely than human interviewers to use acknowledgements and to pivot to higher-level questions. Realizing that a fundamental deficit exists in multi-turn planning and strategic thinking, we develop a realistic simulated environment, incorporating source personas and persuasive elements, in order to facilitate the development of agents with longer-horizon rewards. Our experiments show that while source LLMs mimic human behavior in information sharing, interviewer LLMs struggle with recognizing when questions are answered and engaging persuasively, leading to suboptimal information extraction across model size and capability. These findings underscore the need for enhancing LLMs' grounding and strategic dialogue capabilities.

1 Introduction

Recent research has shown that LLMs struggle to engage in emotional (Shaikh et al., 2024a) or strategic (Wongkamjan et al., 2024) dialogue. For example, Shaikh et al. (2024a) examined LLM-generated responses to dialogues and found less occurrences of "grounding language" (Clark, 1996; Cho and May, 2020), like acknowledgements or affirmations, that humans typically use to foster comfort and trust. This can impede an LLM's ability to serve in a variety of situations: e.g. education (Kasneci et al., 2023), mental health (Carlbring et al., 2023) or conflict resolution (Argyle et al., 2023). However, prior efforts to ameliorate such gaps face limitations: existing large datasets (1k-10k transcripts) are generated via crowdsourcing and are inherently unnatural (Rashkin et al., 2019; Wang et al., 2019; Liu et al., 2021). More natural transcripts, of educational (Caines et al., 2020) or therapeutic environments (Gratch et al., 2014), are difficult to collect due to privacy concerns (Casey, 2004) and are small-scale (100-1k transcripts).

In this work, we directly address this by focusing on an area where real-world and grounding communication is required but plentiful data exist: journalist interviews. Journalistic, or *informational* interviews, are typically conducted between an "interviewer" and a "source", and the goal is to obtain information. Sources are often anxious or unclear (Harcup, 2015), and professional human interviewers are constantly evaluating: (1) Are my questions getting fully addressed? (2) Do I need to alter my communication pattern to more effectively engage or persuade source when necessary (Sedorkin, 2015)? This makes news interviews an ideal setting to develop strategies for grounding, persuasive dialogue. To study how LLMs perform in grounded and persuasive conversations, we start by collecting interviews from two major news sources: National Public Radio (NPR) and Cable News Network (CNN). We clean our dataset, filtering down to 2-person informational interviews (e.g. as opposed to games, questionaires and other formats these news outlets release), resulting in a dataset of 40,000 interviews[1]. Next, we show that news interviews suffer from the same lack of grounding language as has been observed in other dialogue settings (Shaikh et al., 2024a). We perform a discourse analysis, utilizing a schema we develop for interview question-types. We find that, comparing LLM behaviors with human behaviors, significant

[1]We will publicly release this dataset upon acceptance.

distributional differences exist in the kinds of questions asked by LLMs and human interviewers: for example, LLMs are 50% less likely to make acknowledgements, and 30% less likely to pivot back to higher-level questions than humans are.

Motivated by our observations, we develop a realistic simulated environment to serve as a playground for iteratively improving dialogue agents. We processes each of our interview transcripts to extract both (1) the interviewer's high-level objectives and (2) the source's available information. We implemented three difficulty levels: easy, medium, and hard. In the easy level, the interviewer poses questions, and the source responds with relevant informational items, which we track as a reward signal. The intermediate level introduces source-personas (e.g., anxious, clueless, dominating), each with specific limitations on communication patterns and the completeness of their answers. The hardest level adds an element of strategic dialogue: sources will only return more information if they are persuaded in a manner befitting their personas [2]. We carefully design the source's level of persuadability by tracking conversational cues studied in journalism (Harcup, 2015; Sedorkin, 2015).

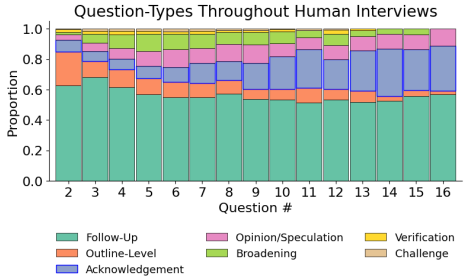
In our setup, both interviewer and source are simulated using LLM agents. We find that the source-LLMs are similar to human players in their information sharing and responsiveness to persuasion. However, interviewer-LLMs struggle to both (1) recognize when questions are answered and (2) actively persuade the source, resulting in suboptimal information extraction compared to human players.

In summary, our contributions are:

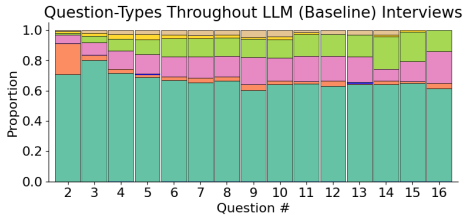
- We collected and processed a dataset of 40,000 two-person informational interviews from NPR and CNN, which we will make publicly available. This dataset addresses the scarcity of large-scale dialogue data necessary for studying grounding communication.

- We performed a detailed discourse analysis comparing LLM-generated dialogues with human interviewers, identifying significant differences in the use of grounding language and question types.

[2]We understand that "being persuaded", "being made comfortable" and "being acknowledged" are all separate forms of grounding, some more active than others. However, we use "persuasion" as a short-hand, as often journalists do need to actively persuade.



(a) **% of Discourse types throughout human interviews**. Human journalists use different discourse roles across the interview, including gradually more Acknowledging statements, increasing from 5% at the start to over 20% by the end.



(b) **% of Discourse types of LLM responses in interviews**. LLMs display an increasing likelihood of asking opinion or broadening questions over the course of an interview and a lower likelihood of returning to outline-level questions.

Figure 1: Comparison of discourse types across interviews (the first turn, usually a greeting, is excluded from analysis). To generate LLM questions, LLM is shown the first $k-1$ turns of a human interview and asked to generate the next question.

- We developed a realistic simulated environment with varying difficulty levels to test and improve dialogue agents. Our findings highlight the challenges LLM interviewers face in recognizing answered questions and engaging in persuasive dialogue.

2 Dataset Processing

In this section, we describe the collection and processing of our dataset of interview transcripts.

2.1 Data Collection

We aggregate, clean and condense multiple publicly available datasets of interview transcripts from National Public Radio (NPR) and Cable News Network (CNN)(Majumder et al., 2020; Zhu et al., 2021) in order to build a high-quality interview dataset. We start with 487,310 transcripts. These transcripts are published records of live interviews conducted between a journalist and sources invited on the program. They provide a rich resource for analyzing natural language interactions in informational interviews.

2.2 Data Filtering for Interview Analysis

We want to focus on one-on-one informational interviews between a journalist and a single source. However, initial examination of the transcripts revealed that many included multiple sources, were formatted as panel discussions, or were not informational in nature (e.g., they included game shows or entertainment segments). To filter the transcripts and retain only those that fit our criteria, we prompted Llama-3.1-70b[3] to classify each transcript based on the number of participants and the nature of the content. The prompts used for filtering are provided in App. C.5. After filtering, we are left with 45,848 interviews meeting our criteria.

After filtering the set of transcripts, we needed to accurately identify which participant was the interviewer vs. the interviewee. We implemented a simple heuristic: we counted each participants' use of question marks. The participant with a higher count was assigned the interviewer role. Manual validation on 50 randomly sampled interviews indicated this method was highly accurate, correctly identifying roles in over 98% of cases.

3 Analysis

In this section, we analyze how humans conduct informational interviews and compare this behavior to that of pretrained LLMs. Our aim is to explore whether LLMs face similar grounding problems in interviews as observed in educational and therapeutic settings (Clark, 1996; Shaikh et al., 2024b).

3.1 Generating Counterfactual Utterances

One way to assess offline how an LLM would behave in an interview setting is to perform a counterfactual simulation, similar to Shaikh et al. (2024b). Specifically, given a human interview, C, consisting of n interviewer-source conversational turns $(q_1, a_1)...(q_n, a_n)$, we feed $t-1$ turns into the LLM context, ($\texttt{LLM}(C_{t-1})$), along with a prompt instructing the LLM to generate the next question. This generates a counterfactual, g_t to what the human would have said in the next turn, h_t. We experimented with different prompting variations: (1) **Baseline**: The LLM is simply asked to produce the next question without additional guidance. (2) **Chain-of-Thought**: The LLM is instructed to reason about the information already provided in the interview, consider what might be left to ask, and then generate the next question. (3) **Outline**: In addition to a chain-of-thought prompt, the LLM is also provided with an outline of the interview goals (extracted as described in Section 4.2) and is told to incorporate this into its chain-of-thought reasoning. We include full prompt examples for all three variations in Appendix C.5. All question-generation experiments were conducted using `Llama-3.1-70b`.

3.2 Evaluating LLM Counterfactuals

To analyze how similar LLM questions, g_t are to human questions, h_t, we perform two analyses: (1) a **Consistency**-based analysis and (2) a **Discourse** based analysis. In the Consistency analysis, we aim to assess how similar g_t is to h_t, whereas in the Discourse analysis, we aim to assess whether g_t plays a similar *function* as h_t does.

Consistency Analysis We design our consistency analysis around an assessment of several lower-level comparison categories, taking inspiration from Saha et al. (2024), who showed that complex evaluation objectives could be broken down into lower-level assessments. Specifically, we analyze the following categories of consistency: *Informational* consistency, where g_t and h_t seeks similar informational objectives; *Motivational*, where g_t and h_t aim to steer the conversation in a similar direction; *Style*, where g_t and h_t are tonal or structural similarities; *Contextual* consistency, where g_t and h_t each are appropriate within the given context; *Discourse* consistency, where g_t and h_t have similar discourse roles. Finally, putting these together, we assess an $Exact$ match in these dimensions. We ask an LLM, `GPT-4o`, to perform this assessment and manually inspect it's outputs and reasoning threads. We confirm that GPT-4o has a high level of accuracy.

Discourse Analysis To facilitate our analysis, we developed a discourse schema to describe the role each question plays in an interview. This schema includes the following elements: *Follow-up Question* (e.g. "Can you tell us more?", *Outline-Level Question* (e.g. "Moving on, can we discuss the next event?"), *Acknowledgement Statement* (e.g., "I see, that sounds like a scary experience."), *Opinion/Speculation* (e.g. "What do you think will happen?"), *Broadening Question* (e.g. "How does this fit into the broader trend?"), *Verification Question* (e.g. " So to confirm...") and *Challenge Question* (e.g. "These dates don't line up."). See Table 5

[3] https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct, a successor to Llama-2 (Touvron et al., 2023). We ran using the vLLM framework (Kwon et al., 2023)

	Exact Match	Info.	Motivation	Style	Discourse	Context
Baseline-LLM	3.9%	4.4%	4.7%	11.9%	36.2%	53.0%
Chain-of-Thought (CoT)	4.5%	3.6%	5.2%	12.8%	37.0%	56.9%
LLM w. Outline	3.7%	3.8%	4.1%	9.6%	36.2%	46.6%
Outline-CoT	3.6%	3.9%	4.3%	8.3%	29.9%	43.1%
Human	8.2%	17.5%	35.4%	40.2%	54.5%	60.3%

Table 1: **Alignment of LLM-Generated Questions with Original Interview questions**. We give an LLM, `Llama-3.1-70b`, the prior $k-1$ turns in an interview and prompt it to ask the next question. We measure the percentage of times this question aligns to the question the human asked across six dimensions: Exact Match (nearly exactly the same as the original utterance), Information (relevant factual content), Motivation (same motivation as the original question), Style (alignment with tone and phrasing), Discourse (structural role within the interview), and Context (incorporation of contextual knowledge). The prompting strategies compared are Baseline-LLM, Chain-of-Thought (CoT), LLM with an Outline, and Outline-CoT; and, we conduct a human baseline trial with a former professional journalist.

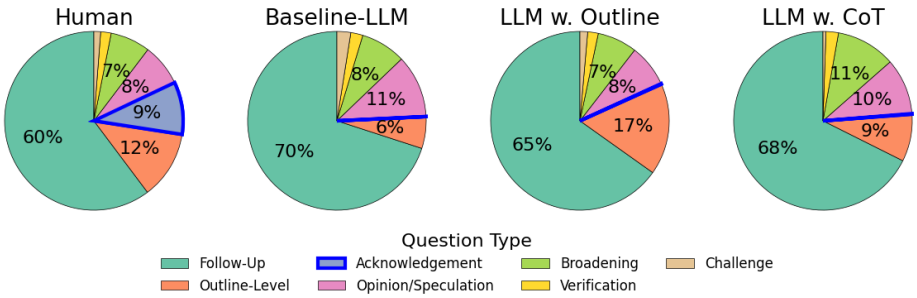


Figure 2: **Distribution of Discourse Roles in Questions, Across Different Prompting Strategies**. We compare the proportions of discourse roles of questions (e.g. "Follow-up", "Acknowledgement", etc.) generated by (a) human journalists, (b) Baseline-LLM (`Llama-3.1-70b`) (c) LLM prompted with an Outline and (d) with Chain-of-Thought (CoT). Acknowledgement statements, which often build empathy, are significantly underrepresented in all LLM prompting approaches, compared to human-generated questions (see appendix for Outline-CoT).

in the Appendix for definitions of each discourse role. In order to generate our discourse schema, we analyzed over 50 interview transcripts along with two journalists. One journalist was a professional with over 8 years of experience in newsrooms, and the second was a journalism undergraduate student. We held three conferencing sessions where we annotated sentences, allowing each sentences to be multiply annotated with discourse roles. We iteratively expanded our schema and collapsed categories until we were not adding new categories. Then, we blindly annotated 10 interviews, achieving a $\kappa = .6$

Given our schema, we then developed a prompt for an LLM to classify discourse roles in sentences. The prompt contains the preceding context of the interview and the target question, asking it to label the questions' discourse role. To validate the LLM's labeling accuracy, we had the professional journalist label 10 additional interviews as ground-truth and scored the LLM's assignments. The LLM scored a .8 f1 score at annotating these categories, indicating that the LLM has reasonably high accuracy in labeling discourse roles.

3.3 Findings

Our analysis yielded several key findings.

Insight #1: Acknowledgement statements are virtually absent from all LLM variations. As shown in Figure 2, grounding gaps exist in journalistic interviewing similar to those observed by (Shaikh et al., 2024a). While human journalistic interviewers tend to make Acknowledgement statements in about 9% of their utterances, all prompting variations that we experimented with made close to zero acknowledging statements. As shown in Appendix **??**, this lack of acknowledgement is paired with a notable lack of mirroring of the source's question-asking; human journalists, on the other hand, bring a degree of character and voice to an interview.

Insight #2: LLMs do not engage in strategic multi-turn questioning . Even in settings where

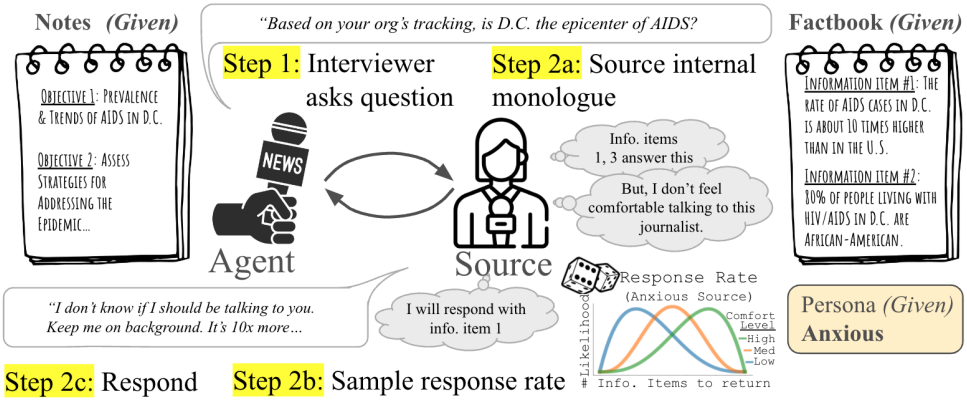


Figure 3: **Walkthrough of the LLM Interviewer-Agent Process**. To set up the interview, the interviewer agent is given a set of high-level objectives, similar to a journalist's pre-interview notes, while the source is given a persona and a set of relevant facts. In each turn, the interviewer asks a question (Step 1). The source determines what information to reveal based on relevance and comfort level (Step 2a). Depending on their comfort, a subset of relevant information is randomly selected for the response (Step 2b). The source then crafts a reply aligned with their persona (Step 2c). The reward given to the interview agent, at the end of k turns, is the number of information items extracted from the source.

LLMs are exposed to interview outlines, they are still undirected in their questions. As shown in Figure 2, LLMs are significantly more likely to ask follow-up questions than humans across all prompting variations. Introducing chain-of-thought and outline variations increased the rate at which the LLM asks outline-level questions. However, the rate remains significantly below human levels. Additionally, they are also more likely to ask either Opinion questions or Broadening questions. In fact, in Figure 1b, we observe that LLMs tend to ask increasing amounts of Opinion Questions and Broadening Questions over time, which humans do not. As can be seen in Table 8, these questions can be vague and open-ended and reflect a lack of planning. Together, these findings suggest an inability to direct an interview in a desired direction and engage in multi-turn planning.

Insight #3: LLMs are capable of understanding context, but fail in all other categories of similarity to humans. Comparing the content and style of LLM interviews to human interviews in Table 1, we note that, overall, LLMs are broadly dissimilar to humans in style, motivation and information-seeking. One area where the LLMs succeed, relatively, is understanding the context of the interview beforehand. This is not a new observation – much recent work, e.g. in dialogue-tracking, has found LLMs to perform well (Ou et al., 2024). The fact that LLMs can hold context over multiple turns and do not drift away from the topic indicates that the LLM might *one day* be able to engage in multi-turn goal-oriented dialogue, given the right reward signals and learning environment.

Taken together, these findings suggest that journalistic dialogue is suitable for studying effective communication patterns, and also highlights significant gap in current language modeling objectives. While LLMs can generate contextually relevant questions, they lack both an emotional and connective drive as well as the strategic planning exhibited by human interviewers.

4 Experimental Setup

In our previous analysis, we observed that LLM counterfactual questions exhibit several shortcomings: they are less likely to acknowledge the interviewee, focus excessively on follow-up questions and lack strategic multi-turn planning. Our goal for the remainder of the paper is to set up a realistic simulated game-environment with a delayed reward signal to train multi-turn agents at scale.

4.1 Game Design Overview

We first introduce our game on a high level, our game design, as illustrated in Figure 3, and then describe our implementation. We draw on two widely used journalism textbooks: *Interviewing: A Guide for Journalists and Writers*, which explains how to conduct effective interviews and deal with reluctant, defensive, or poor-explaining sources (Sedorkin, 2015); and *Journalism: Principles and Practice* which provides guidance on building source trust (Harcup, 2015).

Algorithm 1 Gameplay

Input Interviewer objectives o, Source Informational Items I, Source persona ϕ, K turns
Output Reward R

Initialize: Reward $R \leftarrow 0$, Conversation History $c \leftarrow []$, Used items $u \leftarrow \{\}$
for i $\in 1, ...K$ **do**
▷ **Step 1: Interviewer Question Generation**
  $q =$ `Interviewer`(c, o)
▷ **Step 2: Source's Response Generation**
  $r =$`getRelevantInfoItems`(I, u, q)
  $p =$`getPersuasionLevel`(c)
  $f =$`getItemsToReturn`(r, p)
  $a =$`Source`(q, c, f, p, ϕ)
▷ **Update Variables**
  $u \leftarrow u \cup f, c \leftarrow c \oplus [q, a], R \leftarrow R + |f|$
end for

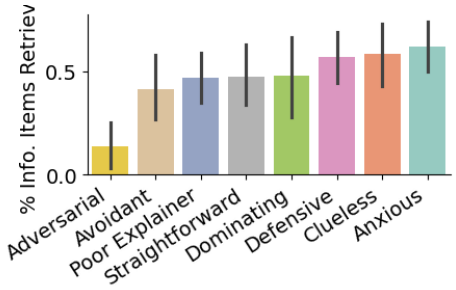
Our gameplay proceeds in a loop, shown in Algorithm 1. The "player" in our game plays the role of an interviewer and is able to ask questions to a source, based on the conversational history and the interview objectives (the `Interviewer()` step). The source is given a set of informational items, and assesses whether any of these items are relevant to the question (the `getRelevantInfo()` step); they then decide how persuaded or comfortable they are based on the conversational history (the `getPersuasionLevel()` step); based on this, we determine the subset of relevant items they return (the `getItemsToReturn()`). They respond with these items. The reward, awarded at the end of the game, is that amount of information items the source disclosed.

4.2 Implementation Details

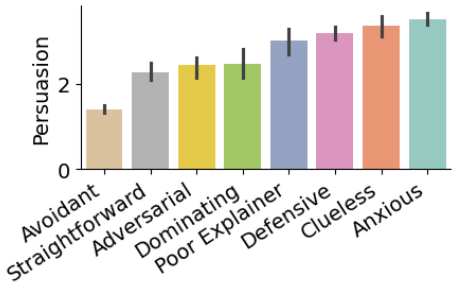
We first start by describing our data processing, and then we will describe the functions introduced in Algorithm 1 in more detail.

Dataset Preparation for Simulation To prepare our dataset for use in the simulated game environment, we performed further processing. To generate the source's information, we group sources responses and used Llama-3.1-70b to summarize them into a set of specific informational items. This represents the knowledge that the source had going into the interview. Manual evaluation confirms that these information items were present in the initial interview, are concise and non-overlapping.

Next, we group interviewer questions and again used Llama-3.1-70b to summarize them into a set of high-level objectives. These objectives represent the key topics or information the journalist aimed to uncover during the interview, reflecting the possible agendas they had prior to the conversation. Manual validation with professional journalists confirms that these outlines reasonably capture what a journalist might prepare before an interview and do not leak information. For examples of both, please see Tables 7 and 6 in the Appendix.



(a) Rewards of `gpt-4o` from playing against sources of different persona types.



(b) Average level of persuasion, from `gpt-4o`, towards the different persona types in our evaluation.

Figure 4: Comparison of `gpt-4o`'s performance across different persona types: extraction rewards and persuasion levels. The Adversarial type is by far the hardest to extract information from, however, it is easier to persuade. This could be because the LLMs are the most thrown off by adversarial sources.

Source Design: Personas and Persuasion Now, we introduce the design of the source. We focus most of our attention on this construction in order to build a robust game environment that accurately mimics human interactions. To make gameplay varied and challenging, we drawing from (Sedorkin, 2015) to design 8 different personas: *Anxious*, *Avoidant*, *Adversarial*, *Defensive*, *Straightforward*, *Poor Explainer*, *Dominating* and *Clueless*. To see detailed descriptions of each persona, as well as example responses that the persona might give, see Table 3. Different personas give us the ability to study how interviewers perform in a wider array of scenarios and are designed to more carefully capture different challenges that journalists face.

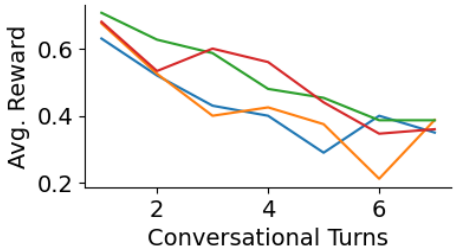
The following functions power internal mechan-

Model	**Hardest** Full Game	**Intermediate** *sans.* Persuasion	**Easiest** *sans.* Info. witholding
gpt-4o-mini	49.3%	47.5%	84.7%
gpt-4o	50.4%	49.8%	84.2%
Llama-3.1-70b	42.6%	45.5%	80.1%
Llama-3.1-8b	42.4%	48.3%	74.9%

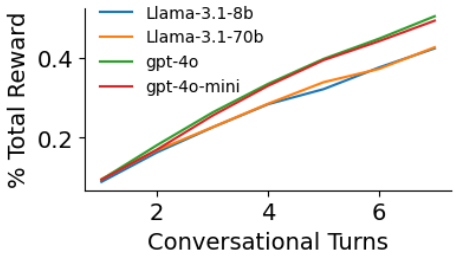
Table 2: **Performance of LLMs as Interviewers, with Ablations** Percentage of information items extracted (Reward %) in each interview by different language models (gpt-4o-mini, gpt-4o, Llama-3.1-70b, and Llama-3.1-8b) across three conditions: (1) **Hardest**: The full game, with information dependent on persuasion and persona (2) **Medium**: an ablation removing the source's responsiveness to persuasion. (3) **Easy**: An ablation removing the random witholding of information (i.e. the source returns all relevant information items at each turn). We observe, perhaps unsurprisingly, that removing the source's ability to withold information (Intermediate $\rightarrow$ Easy) drastically increases the reward % at the end of the game. The removal of persuasion strategies has a smaller effect, with some models showing marginal gains (e.g., Llama-3.1-8b) and others slight losses (e.g., gpt-4o). This indicates that vanilla LLMs are poorly suited to this persuasion task.

ics in our gameplay: `getRelevantInfoItems()`, `getPersuasionLevel()`, `getItemsToReturn()`. `getRelevantInfoItems()` is simply a retrieval function: it takes the interviewer's question and determines which of the sources information items are most relevant. `getPersuasionLevel()` is a function that determines the source's level of comfort or persuasion (on a 1-5 point scale) in the current conversation.

We wished to implement a persuadability component to our gameplay, as this increases the multi-turn strategy an interviewer can play: words spent making a source feel comfortable might be wasteful in a single turn response, but can play a greater roll by eliciting more reward later in the conversation. However, we acknowledge that this is the most challenging part of our setup, as recent research has found that LLMs are poor detectors of when they are being persuaded (Sakurai and Miyao, 2024) and can even unknowingly persuade themselves (Zeng et al., 2024). As a starting point, we draw from (Sedorkin, 2015), and carefully design a series of prompts asking an LLM to rate the persuasiveness of a prior conversation to implement `getPersuasionLevel()`. Persuadability also varies from person to person (Wang et al., 2019; Hirsh et al., 2012), and different source personas, according to (Sedorkin, 2015), are persuaded by different communication patterns. So we also tailor different persuasive phrases for different simulated personas: e.g. *Anxious* sources are distrustful of journalists; they are usually persuaded by phrases like "I will be as fair as possible". For a full list of our persuasion, please see Table 4. With these steps, we can simulate persuadability better and also invite future work on more realistic simulation methods.



(a) Average reward across conversational turns.



(b) Percentage (%) of Reward, by total reward, over time.

Figure 5: Comparison of Rewards over time for language models. For all language models, the reward declines over time, shown above. However, this is not due to interviewer "maxing out" reward, as Total Reward increases nearly linearly across conversational turns.

We confirm the validity of LLM-identified persuasion levels by setting up human trials. In these trials, the human plays as a source. The human receives source information items and is asked questions by the interviewer. After each question, the human rates the their own level of persuasiveness on a 1-5 point scale and we calculate the agreement between the human and the source. After 72 trials, we report $r = .43$ at $p < .0001$, a sta-

tistically significant moderate correlation (if we exclude "Adversarial Sources", our correlation is $r = .68$ at $p < 1e9$, indicating that persuasiveness evaluation in this context is relatively objective.

Finally, based on the assessed persuasion level (1-5) of the conversation, we implement `getItemsToReturn()`. This function takes in all relevant information items and randomly draws from a Beta(α, β) distribution to determine what percentage of relevant information items to return. We implement this persuasion-dependent sampling by choosing 5 parameterizations per persona $(\alpha_{p,1}, \beta_{p,1}), (\alpha_{p,2}, \beta_{p,2}), (\alpha_{p,3}, \beta_{p,3})$.... As can be seen in Figure 3, we choose these parameterizations such that the more persuaded a source is, the more left-skewed the distribution is. Each persona has a slightly different parameterization, reflecting how some personas need less persuasion (e.g. "Dominant") while others do not drastically change how much information they return even with more persuasion (e.g. poor explainer). See Figure 6 in the Appendix for the Beta distributions chosen for each source.

Source and Interviewer Responses To implement the conversational components of our game, we stick with a minimalist setup, as LLMs have been well-observed to be proficient at basic dialogue given sufficient structure (Ou et al., 2024). For `Source()`, we provide the source with the information items to return, a description of it's persona, and it's prior assessed comfort/persuasion level. For `Interviewer()`, we intentionally implement a minimalist interviewer agent and intend to leave more involved designs for future work.

5 Results and Discussion

We ran our simulation for 200 interviews with 4 models as the interviewer: `gpt-4o`, `gpt-4o-mini`, `Llama-3.1-70b` and `Llama-3.1-8b`. For a source LLM, we used `gpt-4o` across all personas. Table 2 compares the performance of LLMs across three conditions: the full game, a version without persuasion, and a version where sources do not withhold information. In the full game, where sources' responsiveness depends on persuasion and persona, the gpt-4o model performs the best, at 50.4%. However, when persuasion was removed, performance only marginally improved across all models (e.g., Llama-3.1-70b reached 45.5%, while gpt-4o remained stable at 49.8%), indicating that other aspects of the game (i.e. inferring which information the source has witheld) also pose a challenge. In the easiest condition, where no information withholding occurred, all models performed significantly better, with reward percentages reaching over 80%, showing that withholding strategies are a major obstacle for current LLMs.

Figure 4a highlights the performance of gpt-4o across different source personas. The model achieved the highest information extraction from straightforward personas, while adversarial and defensive personas proved the most challenging. Despite being harder to extract information from, adversarial sources were easier to persuade (Figure 4b), suggesting that the LLM models are more responsive to engagement when sources display hostility rather than indifference or avoidance.

Figure 5a explores how the reward (information extraction) changes over the course of an interview. The results show a declining trend in reward per conversational turn, suggesting that LLMs struggle to maintain effective information extraction as the conversation progresses. This may be due to difficulties in adjusting their questioning strategies over time. However, the total reward accumulated over time (Figure 5b) increased almost linearly, showing that the LLMs continue to extract information, albeit at a slower rate. Together, these findings highlight the limitations of current LLMs in engaging with persuasive and strategic multi-turn interviews. While larger models like gpt-4o outperform smaller ones, they still exhibit significant gaps in persuasion and adaptive questioning, particularly when dealing with difficult personas.

6 Conclusion

Our game-based evaluation reveals that while LLMs struggle to extract information from a variety of personas. Interestingly, particular personas, like adversarial or avoidant sources, are noticably more difficult for the LLM to deal with. Finally, we see only marginal gains with large modeling sizes, suggesting that scale alone may not make meaningful progress. Our work builds off other work that has observed this (Shaikh et al., 2024a), offering an opportunity to train LLMs in multi-turn dialogue with delayed rewards. We have created as human-like a game environment as possible. We look forward to future work building off of this groundwork.

7 Limitations

7.1 Privacy and Ethical Considerations

All data used in this study are publicly available and do not contain personally identifiable information beyond what has been already made public by the news organizations. We adhere to ethical guidelines for data use and ensure that our processing respects the rights and privacy of individuals involved as well as the news organizations that collected this data. Since the dataset we create is derived from interviews that have already been published in academic settings, we believe we are infringing upon the copyright of the news organizations this data originally belonged it.

Aside from ownership questions, there are still inherent risks in the use of real-world interview data. Some interviews might involve sensitive topics, and the ethical implications of using such data for model evaluation warrant careful consideration.

7.2 Reproducibility

All experiments are conducted using publicly available models and datasets. The code and data necessary to reproduce our results will be made available upon submission. Part of our simulation does rely on high-performing language models and to serve this we implemented used `gpt-4o`. This brings us into territory where we are inherently not reproducible, as closed models can be changed without notice. However, we believe we are not out of the norm in the academic community in our usage.

7.3 Simulated Environment Limitations and Risks

The simulated game-playing environment used to evaluate the LLM agents is a simplification of real-world interviewing processes. We might be inducing a bias in agents. We also might be opening up a sandbox for potential dual-use. The design of our game, to extract information from sources, might one day be used to persuade users to divulge sensitive information.

8 References

References

Lisa P Argyle, Ethan Busby, Joshua Gubler, Chris Bail, Thomas Howe, Christopher Rytting, and David Wingate. 2023. Ai chat assistants can improve conversations about divisive topics. *arXiv preprint arXiv:2302.07268*.

Andrew Caines, Helen Yannakoudakis, Helena Edmondson, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2020. The teacher-student chatroom corpus. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 10–20.

Per Carlbring, Heather Hadjistavropoulos, Annet Kleiboer, and Gerhard Andersson. 2023. A new era in internet interventions: The advent of chat-gpt and ai-assisted therapist guidance. *Internet Interventions*, 32.

Dympna Casey. 2004. Challenges of collecting data in the clinical setting. *NT Research*, 9(2):131–141.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tony Harcup. 2015. *Journalism: Principles and Practice*, 3rd edition. SAGE Publications, London, UK.

Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. 2012. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6):578–581.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3469–3483.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: A large-scale open-source corpus of media dialog. *arXiv preprint arXiv:2004.03090*.

Jiao Ou, Junda Lu, Che Liu, Yihong Tang, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Dialogbench: Evaluating llms as human-like dialogue systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6137–6170.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8345–8363.

Hiromasa Sakurai and Yusuke Miyao. 2024. Evaluating intention detection capability of large language models in persuasive dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1635–1657, Bangkok, Thailand. Association for Computational Linguistics.

Gail Sedorkin. 2015. *Interviewing: A Guide for Journalists and Writers*, 4th edition. Allen & Unwin.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024a. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024b. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Wichayaporn Wongkamjan, Feng Gu, Yanze Wang, Ulf Hermjakob, Jonathan May, Brandon M Stewart, Jonathan K Kummerfeld, Denis Peskoff, and Jordan Lee Boyd-Graber. 2024. More victories, less cooperation: Assessing cicero's diplomacy play. *arXiv preprint arXiv:2406.04643*.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.

A Additional Details on Source Prompts

B Discourse Definitions

We give discourse definitions for our 8 discourse categories in Table 5. We developed these definitions between 3 annotators, one of which was a former professional journalist, another was a journalist undergraduate student and the third was a computer science undergraduate student. We conferenced 3 times, examining over 50 interviews, sorting questions into categories, and expanding these categories until we were reliably able to label new questions. We calculated an inter-annotator agreement of κ = .6 between the annotators on a shared set of 10 interviews. Then, we used an LLM, `Llama-3.1-70b` to label discourse on the entire interview. The former journalist manually evaluated the LLM's performance during a blind trial and had an agreement of $\kappa = .8$ with the LLM.

C Data Preprocessing

C.1 Data Used

- `utterances-2sp.csv` from the NPR-Media dataset (Majumder et al., 2020)
- `episodes.csv` from the NPR-Media dataset (Majumder et al., 2020)
- `news_dialogue.json` from the MediaSum dataset (Zhu et al., 2021)

C.2 Initial Sizes of the Data

- There are 1,240,112 rows and 7 columns in `utterances-2sp.csv`.
- There are 105,848 rows and 4 columns in `episodes.csv`.
- There are 3,199,858 rows and 4 columns in `utterances.csv`.
- There are 23,714 transcripts in the NPR-Media dataset (from `utterances-2sp.csv`).
- There are 463,596 transcripts in the MediaSum dataset.

C.3 Process

C.3.1 NPR-Media

- Began with combining the extttepisodes.csv with the `utterances-2sp.csv` to add more information about the episode (title, date, etc).
- Filtered out based on keywords: ["Sunday Puzzle", "Traffic", "Puzzle", "Advertisement", "Sponsor", "Commentary"]
  - 37 interviews were filtered out and reduced to 23,676 transcripts.
- Helper functions used:
  - `count_unique_episodes(df)` allows the user to count the number of unique episodes within a dataset.
  - `filter_interviews(merged_df)` allows the user to filter out the dataset using certain keywords.
  - `find_removed_episodes(df_before, df_after)` allows the user to identify the episodes that were removed from the dataset.
  - `print_episode(df, episode_number)` prints a specified episode from a specified dataset.
  - `print_episode_pretty(df, episode_number)` prints a specified episode from a specified dataset with a readable format.
  - `find_removed_episodes(df_before, df_after)` returns a list of the episodes that were removed in a filtering step.
- Convert dataset to the MediaSum format for easy prompt processing.
- Downloaded dataset (grouped and ungrouped) as JSON and CSV.

MediaSum

- Began with deleting interviews from MediaSum that are already in NPR-Media.
  - 8 interviews were filtered out and reduced to 463,588 transcripts.
- Filtered out episodes that had more than two unique speakers in the middle 70% of the transcript.
  - Reduced to 66,978 transcripts.
  - 396,610 interviews were filtered out and reduced to 66,978 transcripts.
- Filtered out episodes that were too short.
  - 19,059 interviews were filtered out and reduced to 47,919 transcripts.

Persona	Description	Example Responses
Anxious	Unsure if they should be doing the interview, often expresses doubt.	"I'm not sure if I should be saying this, I should speak to my manager."
Avoidant	Brief, deflects questions, avoids detail, and changes subjects.	"Actually, one of the main issues was the supply chain, but we've sorted that out."
Adversarial	Hostile, challenges the interviewer, provides confrontational replies.	"Maybe if you did your job well, you'd understand the data. I'm not here to educate you."
Defensive	Protects reputation, feels criticized, gives overly detailed explanations.	"One area where costs increased was in material prices, which were out of our control."
Straightforward	Clear, direct, and willing to provide detailed information.	"Additionally, we ran out funding midway through the project."
Poor Explainer	Struggles to explain clearly, rambles, or provides convoluted answers.	"Uh, well, I guess the supply chain was part of it, but, uh, that's only one part of the story..."
Dominating	Controls the conversation, gives lengthy or off-topic answers.	"Costs were high, but at my suggestion we brought in the best experts worldwide."
Clueless	Confused and uncertain, often unsure of the topic.	"Oh, right, the delays... yeah, maybe it was the, uh, supply issues? I'm not too sure..."

Table 3: Source Personas that we created, with Descriptions and Example Responses

- Helper functions used:
  - `print_row_by_title(df, title)` allows users to print episode using title.
  - `print_row_by_id(df, id)` allows users to print episode using id.
  - `filter_episodes_2sp(df)` filters out transcripts with more than 2 unique speakers in the middle 70%.
  - `filter_by_utt_length(df)` filters out transcripts with 10 or fewer strings in extttutt.
  - `find_removed_episodes(df_before, df_after)` lets users see which episodes were removed.
- Downloaded dataset as JSON and CSV.

C.4 Final Sizes of the Data

- There are 23,676 transcripts in the NPR-Media dataset (from extttutterances-2sp.csv).
- There are 47,919 transcripts in the MediaSum dataset.
- There are 71,598 transcripts in the combined dataset.
- There are 45,848 transcripts in the final dataset.
- Our dataset started at 487,310 transcripts and now has 45,848 transcripts.

C.5 LLM Preprocessing Prompt

Prompt to filter out transcripts that were not informational interviews `Analyze this interview transcript that is in the form of a dialogue: (dialogue) By reading through the dialogue, identify if this transcript is an informational interview between 2 people. Look for questions and make sure this is an interview, not a Q&A game. The interviewer should be asking questions, not engaging in a back-and-forth conversation with the interviewee. After analyzing, your final answer of just 'YES' or 'NO' should be in brackets.`

C.6 Examples of Interviews

NPR-85

FARAI CHIDEYA, host: Tony, I guess there will always be some kind of history made every day.

TONY COX, host: You know, some of it good. Some of it, not so good.

Persona	Persuasion Description	Persuasion Examples
Anxious	Responds well to empathetic, reassuring, and patient conversations. Encouraging, non-threatening language builds comfort.	"I will be as fair as possible.", "I appreciate your honesty.", "If you're not comfortable now, I can come back later."
Avoidant	Prefers non-obtrusive small talk, short questions, and space to respond. Open-ended, light prompts work well.	"And that happened when?", "I imagine there's more to the story.", "Ah I see."
Adversarial	Responds to thorough research, persistence, and fact-based questions. Repeated questioning elicits responses.	"Our records indicate...", "Just to be clear, are you saying...?", "Earlier you stated..."
Defensive	Engages with non-confrontational and validating conversations. Neutral language reduces defensiveness.	"I see why you made that choice.", "We can work together.", "It's understandable."
Straightforward	Prefers direct and transparent conversations. Efficiency and brevity are key.	"Let's get to the solution.", "What were the key points, in your view?"
Poor Explainer	Responds well to structured, patient conversations. Simple clarifying questions and validation help communication.	"Explain that part again in smaller steps.", "I understand, keep going.", "Take your time."
Dominating	Engages when their expertise is acknowledged. Validation and offering control builds rapport.	"I'd love your take.", "You have experience, what do you suggest?", "Your insights are valuable."
Clueless	Guided, simple questions with firm direction are effective. Breaking down complex topics increases confidence.	"Tell me what you're thinking.", "It's okay to be unsure.", "Start with something simple."

Table 4: Persuasion techniques that we compiled for different sources types. These manners and styles of speaking were informed by examples given in Harcup (2015) and Sedorkin (2015) that sources with different personality types find the most persuasive.

FARAI CHIDEYA, host: And while some of it is well-publicized, sometimes, notable history goes under the radar.

TONY COX, host: Now, that's true.

FARAI CHIDEYA, host: I'm thinking of your interview with Mable John.

TONY COX, host: Oh, yeah. Now, this is a woman with an interesting past.

Ms. MABLE JOHN (Singer): (Singing) My name is Mable and don't you think I ain't able. TONY COX, host: The 77-year-old Louisiana native has been a top R&B singer, a successful novelist, a pastor, an activist and a movie actor, and I found out that Mable John is full of stories like the one about the time she met record mogul Berry Gordy before Motown was even Motown.

Ms. MABLE JOHN (Singer): (Singing) That you're leaving.

Ms. MABLE JOHN (Singer): How I met Berry? That was at a barber shop on (unintelligible) that was near the fine show bar, and at that time men were wearing process. Process is the (unintelligible). And I was dating a guy that was one of those process operators in the Chesterfield lounge and barbershop, and Berry was coming and getting his hair done. I was coaching choirs for my church. And my boyfriend introduced me to Berry Gordy because Berry said he was a songwriter and he was going to have a lot of people recording his songs. And my boyfriend said you need to stop doing all of this work for the church free, and that Berry Gordy do something with you so you can get paid. So he introduced me to Berry Gordy.

TONY COX, host: Now, tell us the story. We're going to skip around a little bit.

Ms. MABLE JOHN (Singer): Okay.

TONY COX, host: When you and Berry Gordy connected, as Motown was just becoming a com-

Discourse Role	Definition
Starting/Ending Remarks	Initiates or concludes the interview. Often not in the form of a question.
Acknowledgement Statement	Affirms the interviewee, often by explicitly recognizing their previous response. Builds rapport, demonstrates active listening. Typically induces the source to engage in greater openness.
Follow-Up Question	Digs deeper into a topic, seeks elaboration, or re-phrases a previous question to keep the discussion focused.
Verification Question	Confirms the accuracy of a statement, fact, event or observation or assumption.
Topic-Transition Question	Shifts the conversation to a new subject, usually an outline-level goal that the journalist prepared before the interview.
Opinion/Speculation Question	Solicits the interviewee's views or predictions, revealing biases or insights.
Challenge Question	Tests the interviewee's position, argument, or credibility, often provoking thought or debate.
Broadening Question	Expands the scope of discussion, encouraging the interviewee to consider broader contexts or new perspectives.

Table 5: Discourse types in informational interviews. We developed these definitions manually between three annotators by examining 50 different interviews.

Processed Outline
Source biography: Senior news analyst with expertise in politics and history.
Interview context: A president's final year in office and potential changes in policy.
Objective 1: Presidential legacy
Objective 2: Foreign policy shifts
Objective 3: Domestic policy changes
Objective 4: Potential surprises

Table 6: Interview Outline Objectives given to an interviewing game playing agent.

pany, a record company, you are the first female to record on a label, the Tamla label.

Ms. MABLE JOHN (Singer): Yes.

TONY COX, host: Before Motown.

Ms. MABLE JOHN (Singer): The first single female artist, because Claudette Robinson was a part of what become the Miracles, and he was managing them along with me.

TONY COX, host: Right.

Ms. MABLE JOHN (Singer): So I was the first single female artist to be signed to Tamla, which is a part of the Motown family.

TONY COX, host: When you think about that now, how do you feel about looking at that as a historic moment?

Ms. MABLE JOHN (Singer): No one could have bought that time. God had to give it to me.

Ms. MABLE JOHN (Singer): (Singing) Hey. Hey.

TONY COX, host: I understand that you were rehearsing one day and these three young girls came in and interrupted your rehearsal.

Ms. MABLE JOHN (Singer): The girls that we know now as The Supremes. They came into a rehearsal that I was doing with Berry Gordy because he played also for me, played piano for me. We were there rehearsing and these girls came in and I didn't quite remember everything that was said that day because it's been so long. But Mary Wilson of The Supremes, remembered when she was writing her book to say that when she first walked into Motown, the three of them walked in and my question to Berry Gordy was, why are they walking in on my rehearsal, because all of our rehearsals were private.

Ms. MABLE JOHN (Singer): (Singing) It takes a more than 'em flashy old money and I wink from the corner of your eye. I don't want no big line calls, (unintelligible) caviar. Oh no, true love baby can be found 'cause you take a look around. TONY COX, host: Talking about faith. Your career at

Information Items
Row 1: Information item 1: The economy is growing above trend pace, with job growth of 150,000-200,000 a month, which is higher than what's sustainable in the long run. **Row 2:** Information item 2: The Fed will likely continue to raise interest rates, as the economy is growing and financial conditions are still very accommodative. **Row 3:** Information item 3: The neutral rate is probably higher than where we are right now, but it's not a precise number, and the Fed needs to curtail monetary policy further. **Row 4:** Information item 4: The dot-plot is just a forecast and should not be taken as a commitment; it's subject to change as new information becomes available. **Row 5:** Information item 5: The stock market will likely face tougher going in 2019, with slower earnings growth and higher interest rates. **Row 6:** Information item 6: The economy's capacity to continue growing is a concern, as there aren't enough workers to sustain above-trend growth for more than another year or so.

Table 7: Information Items for a source in our game-playing environment, extracted from an interview featuring Bill Dudley, Former President of the New York Federal Reserve

LLM-Generated Counterfactual Questions
What do you think about the changing dynamics of your neighborhood and how it affects your sense of community and belonging?
What specific factors do you think are contributing to the increasing rates of HIV/AIDS among African-American women in Washington D.C.?
How do parents with HIV/AIDS typically cope with the fear of not being there for their children, and what are the emotional and psychological implications of this fear on their mental health?
What specific steps are you taking to mitigate the impact of the Salmonella outbreak on your business, and do you think this will have a lasting effect on the tomato industry as a whole?
What about the potential impact of this rate drop on the overall housing market, and do you think it could lead to a rebound in housing prices or a continued decline?
What are some practical steps parents can take to help their teenagers prepare for the job market and make the most of their summer?

Table 8: List of Interview Questions generated in a counterfactual setting by LLM interviewer. The questions are generated after observing the previous t human conversational turns.

Motown never really took off, and after some few years, you decided to go to Memphis, where you joined the Stax label and hooked up with Porter and Isaac Hayes. And then, it was long after that that you had a million seller.

Ms. MABLE JOHN (Singer): Right. Well, Motown, Berry Gordy, they were all along with God and my parents a part of my future. So Motown was my beginning. It was one that was different from everywhere else I've ever been. But I think it was a necessary one to make the transition for me from Motown to Stax.

TONY COX, host: Now, your big song at Stax, one of your - the biggest of your songs was…

Ms. MABLE JOHN (Singer): The biggest of all songs.

TONY COX, host: "Your Good Thing is About to End."

Ms. MABLE JOHN (Singer): "…Is About to End." Right. Right.

TONY COX, host: See, I'm old enough to have remembered that song.

Ms. MABLE JOHN (Singer): Well, that's good. That makes me feel you don't have to be very old to remember that.

Ms. MABLE JOHN (Singer): (Singing) I don't have to beg you to hold me 'cause somebody else will. You don't have to love me when I want it, 'cause somebody else will.

Ms. MABLE JOHN (Singer): It was a story

that I needed to tell because of a bad marriage. And at Stax, they would allow you to be yourself. Everybody participated in whatever success you're going to have, everybody, including the drummer.

TONY COX, host: Really? Tell me about your family. And I'm switching to that for a reason because you were one of 10 children, right?

Ms. MABLE JOHN (Singer): The oldest...

TONY COX, host: The oldest of 10.

Ms. MABLE JOHN (Singer): ...of 10 children.

TONY COX, host: And you happen to have a little brother, a baby brother who was a big time performer, Little Willie John.

Ms. MABLE JOHN (Singer): Yes. Little Willie John. William Edward John. Now, when I got with Willy that was another education.

Ms. MABLE JOHN (Singer): Because he said my name is Little Willie John. It might be William Edward John to you, and you're my sister and I love you. But if you're not good, I'm going to send you home.

TONY COX, host: Obviously, you are good.

Ms. MABLE JOHN (Singer): Well, he let me stay.

Ms. MABLE JOHN (Singer): (Singing) You have all the love that I've got. Even ice melts to water and gets hot. Look out, your good thing is about to come to an end. Your real good thing...

TONY COX, host: You were the leader of the Raelettes for a dozen years.

Ms. MABLE JOHN (Singer): Yes.

TONY COX, host: Traveling all over with and without Ray Charles.

Ms. MABLE JOHN (Singer): With and without Ray Charles. Yes.

TONY COX, host: In the movie, "Ray," I had looked in the credits to see if there were someone who played you...

Ms. MABLE JOHN (Singer): No.

TONY COX, host: ...since you have been a Raelette for so long, and I saw that there wasn't one.

Ms. MABLE JOHN (Singer): No.

TONY COX, host: And is there a reason for that?

Ms. MABLE JOHN (Singer): Well, it was the years before I came.

TONY COX, host: Okay.

Ms. MABLE JOHN (Singer): And I tell everybody that asks me, the best of his life were the years after the movie. When I came to work with him, he sat me down and told me all about his beginning, told me all about things that ticks him off and things that excite him, what he was looking for and how he wanted it. And I knew that being with him would finish me in this industry...

TONY COX, host: Now, when he...

Ms. MABLE JOHN (Singer): ...because he was at the top - complete me.

TONY COX, host: Okay.

Ms. MABLE JOHN (Singer): So that I could work for any audience, sing any kind of songs. Remember now, at the beginning I thought I could only sing gospel. With Berry Gordy, I found out I could sing the blues. I went to Stax and I find out I could sing love songs. I got with Ray Charles and we sang country - everything. And we could play to any audience. I wanted to sing what was in my heart to everybody that loves music, and Ray Charles was the place for me to be, to do that.

TONY COX, host: So the Raelettes - would you say that was the highlight of your career?

Ms. MABLE JOHN (Singer): It was a highlight. It was highlight because I learned things about myself, about my career, about the industry. I was able to set up my own publishing companies and production companies because of the knowledge that I gained with and from Ray Charles.

TONY COX, host: And after all of that, Mable John, your career did not stop. It has gone on into movies, into - you've written a couple of novels.

Ms. MABLE JOHN (Singer): Excuse me. I just finished the third.

TONY COX, host: Oh, number three. You've done three novels. You're a minister.

Ms. MABLE JOHN (Singer): Yes.

TONY COX, host: And you started a church.

Ms. MABLE JOHN (Singer): Yes.

TONY COX, host: And you help the homeless.

Ms. MABLE JOHN (Singer): Yes.

TONY COX, host: And you're a grandmother.

Ms. MABLE JOHN (Singer): A great-grandmother.

TONY COX, host: And a great-grandmother. How is it possible for one person to do all of those things and to do them as successfully as you have?

Ms. MABLE JOHN (Singer): It's all God. Some days, when people are telling me how busy I am. And when I sit down to think about it, I get tired.

Ms. MABLE JOHN (Singer): So I don't. I don't go there. I just get up every morning and I thank God for the activity of that day. And I have to thank a woman that's no longer with us, Ms. Billie

Holiday, because that's the voice that I hear in my ear still to this day. I worked with her two weeks before she passed. And she said to me, Honey - because I was frightened out of my wits - you can make it if you remember. Always know when you have done or given enough. Not to be afraid and have guts enough to say I quit.

Ms. MABLE JOHN (Singer): (Singing) Even ice melts to water and gets hot…

TONY COX, host: It's so nice talking with you. Thank you for coming in.

Ms. MABLE JOHN (Singer): I thank you.

Ms. MABLE JOHN (Singer): (Singing) Your good thing is about to come to an end. Your real good thing…

FARAI CHIDEYA, host: That was NPR's Tony Cox with singer, author and actor Mable John. Look for Mable John in the upcoming John Sayles film, "Honeydripper."

Ms. MABLE JOHN (Singer): (Singing) Getting myself back together.

FARAI CHIDEYA, host: That's our show for today, and thank you sharing your time with us. To listen to the show or subscribe to our podcast, visit our Web site, nprnewsandnotes.org. No spaces, just nprnewsandnotes.org.

FARAI CHIDEYA, host: To join the conversation or sign up for our newsletter, visit our blog at nprnewsandviews.org. NEWS & NOTES was created by NPR News and the African-American Public Radio consortium. Tomorrow, a reporter shares Donda West's last interview.

FARAI CHIDEYA, host: I'm Farai Chideya. This is NEWS NOTES.

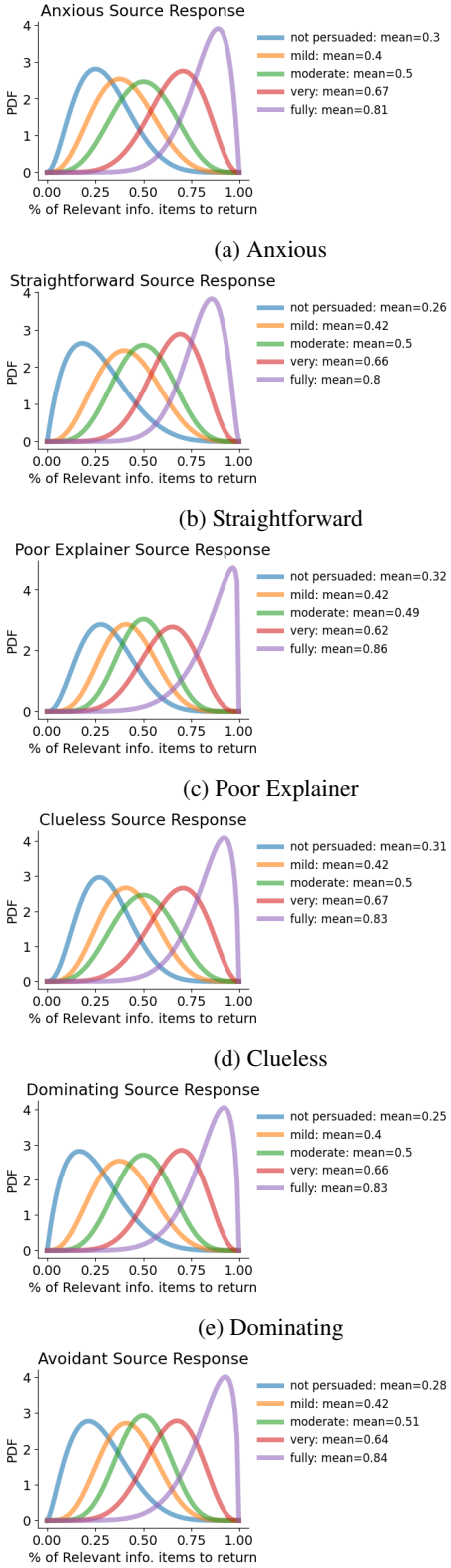


Figure 6: Beta Distributions for Various Interview Personas