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

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in generating coherent text but often struggle with grounding language and strategic dialogue. To address this gap, 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, 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' 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 fewer 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 these limitations by focusing on an area where 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 human interviewers are constantly evaluating: (1) Are my questions getting fully addressed? (2) Do I need to more effectively engage or persuade a source (Sedorkin, 2015)? This makes news interviews an ideal setting to observe grounding dialogues.

To study how LLMs perform in journalistic contexts, we start by collecting interview transcripts from two major news sources: National Public Radio (NPR) and Cable News Network (CNN), filtering to over 40,000 dyadic informational interviews.Next, we show that LLMs in news interview settings suffer from the same lack of grounding as in other dialogue settings (Shaikh et al., 2024a). We find that significant discourse 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 to higher-level questions.

Next, we turn to a more fundamental question: *why grounding?* According to Cialdini (2009), grounding exists for a purpose: to influence an outcome (e.g. in therapeutic environments, the grounded patient is more open and makes more progress (Bohart, 1999); in educational environments, the comfortable student is more engaged and learns more (Brown, 1994); these are all usually borne out over many conversational turns). Observing the *effects* of grounding language in terms of objective outcomes might be a more effec-

tive way to reason and ultimately train empathetic agents and improve long-horizon strategic dialogue. Motivated by these observations, we develop a realistic game environment to serve as a playground: in this simulation, LLMs play the role of the interviewer and the source. The goal for the interviewer is to *obtain the maximal amount of information from the source in a limited number of questions*.

In order to induce the need for grounding communication, we design different personas for sources (e.g., anxious, clueless, dominating), each with different communication patterns. We also add a responsiveness to strategic dialogue: sources will only return information if they are persuaded in a manner befitting their personas[1] (Harcup, 2015; Sedorkin, 2015). We find that our environment is realistic: source-LLMs correlate significantly with humans in their ability to identify persuasion ($r = .43, p < .0001$). However, interviewer-LLMs struggle to both recognize when questions are answered and actively persuade the source, resulting in suboptimal information extraction.

In summary, our contributions are:

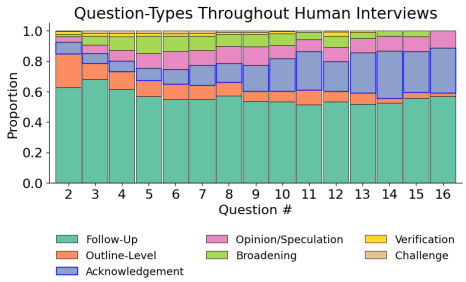
- We release a high-quality dataset of 40,000 two-person informational interviews from NPR and CNN. This dataset addresses the scarcity of large-scale dialogue data necessary for studying grounding communication.
- We perform a detailed discourse analysis comparing LLM-generated dialogues with human interviewers, identifying significant differences in the use of grounding language and question types.
- We develop a game environment to test and improve dialogue agents in informational interviews, which we call *NewsInterview*. Our findings indicate this is a realistic setting but highlight the challenges LLM interviewers face in engaging in persuasive dialogue.

2 Dataset Processing

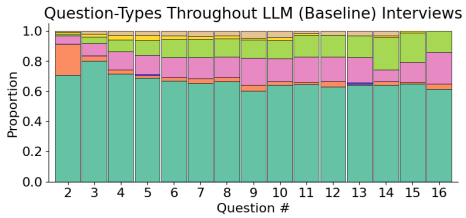
2.1 Data Collection

We aggregate, clean and condense multiple publicly available datasets of interview transcripts from National Public Radio (NPR) and Cable News Network (CNN) in order to build a high-quality interview dataset of 45k source-to-interview 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.

[1]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 encompassing all categories.



(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 (turn 1, usually a greeting, is excluded). LLM is shown the first $k-1$ turns of a human interview and asked to generate the next question.

2.2 Data Filtering for Interview Analysis

We want to focus on one-on-one informational interviews between a journalist and a single source. We start with 487,310 transcripts collected by (Majumder et al., 2020; Zhu et al., 2021). However, initial examination of the transcripts reveals many of them to be low-quality: many include multiple sources, are formatted as panel discussions, or are not informational in nature (e.g., they included game shows). To filter the transcripts and retain only those that fit our criteria, we prompt Llama-3.1-70b[2] 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. F.5. After filtering, 45,848 interviews remain. Finally, the original transcripts do not distinguish which participant was the interviewer vs. the interviewee. So, we count each participant's use of question marks: the participant with more is labeled the interviewer.[3]

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

[3]Manual validation on 50 interviews showed this method correctly identified roles in $> 98\%$ of cases.

	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 a question asked by a human at the same point in the interview 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.

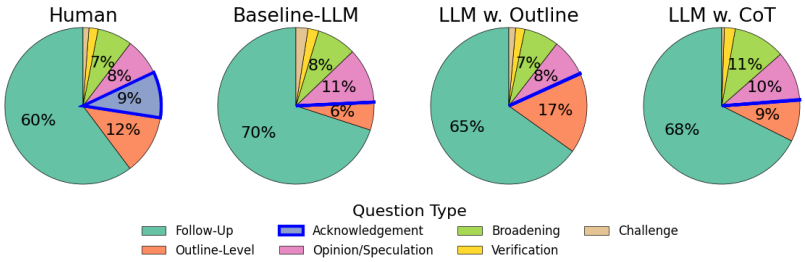


Figure 2: Distribution of Discourse Roles in Questions, Across Different Prompting Strategies. We compare the proportions of discourse roles of questions (e.g. "Follow-up", "Acknowledgement", etc.) generated by (a) human journalists, (b) Baseline-LLM (`Llama-3.1-70b`) (c) LLM prompted with an Outline 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).

3 Analysis

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

3.1 Generating Counterfactual Utterances

One way to assess how an LLM would behave in an interview setting offline is to perform a counterfactual simulation (Shaikh et al., 2024b). Specifically, given a human interview consisting of n interviewer-source conversational turns $(q_1, a_1)...(q_n, a_n)$, we feed $t-1$ turns into the LLM 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, h_t. We experiment with different variations: (1) Baseline: The LLM is simply asked to produce the next question. (2) Chain-of-Thought (CoT): 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: the LLM is provided with an outline of the interview goals (described in Section 4.2) to incorporate into CoT reasoning.[4]

3.2 Evaluating LLM Counterfactuals

To analyze how similar LLM questions, g_t are to human questions, h_t, we perform two analyses:

Consistency Analysis: We aim to assess how similar g_t is to h_t across different comparison categories Saha et al. (2024), specifically: *Informational* consistency (i.e. g_t and h_t seek similar informational objectives); *Motivational*, (i.e. similar outcomes); *Style*, (i.e. similar tone); *Contextual* consistency (i.e. similar appropriateness given the context); *Discourse* consistency (i.e. similar purposes in the overall conversation). Putting these together, we assess an $Exact$ match. We ask an LLM, `GPT-4o`, to perform this assessment and manually inspect its outputs and reasoning threads.

Discourse Analysis: We aim to assess whether g_t plays a similar *function* as h_t does. We develop a schema to describe the role of each question[5]. This schema includes the following elements:

[4]We include full prompt examples for all three variations in Appendix F.5. All question-generation experiments are conducted using `Llama-3.1-70b`.

[5]To generate our discourse schema, we asked two journalists to analyze 50 interview transcripts. One had 8 years of experience in newsrooms, the other was an undergraduate student studying journalism. We held three conferencing sessions to develop the schema. Then, we blindly annotated 10 interviews, achieving a $\kappa = .6$. Given our schema, we then asked an LLM to classify discourse roles in sentences. The prompt contains the interview context, $(q_1, a_1)...(q_{t-1}, a_{t-1})$, and

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 scary."), *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 in the Appendix for definitions of each discourse role.

3.3 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. This lack of acknowledgement is paired with not mirroring the source's speaking style; human journalists, as shown in Appendix F.6, bring character and voice.

Insight #2: LLMs do not engage in strategic multi-turn questioning. Even in settings where 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 increases 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 shown in Table 8, these questions can be vague and open-ended. 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 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 preserve context over multiple turns and do not drift away from the topic indicates that models 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 highlight significant gaps 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.

current question q_t. 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.

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

4 NewsInterview: An Interview Game

As shown, LLM counterfactual questions exhibit several shortcomings: they are less likely to acknowledge the interviewee and focus excessively on follow-up questions. Both of these shortcomings point to a lack of strategic multi-turn planning? In human dialogue, grounding exists for long-term strategic purposes (Cialdini, 2009), yet there currently exists no way to way to obtain these kinds of long-term rewards during LLM training. Motivated by this insight, our goal for the remainder of the paper is to create *and validate* a realistic game-environment with a delayed reward signal. *We leave to future work utilization of this framework for improving strategic dialogue.*

4.1 Game Design Overview

We first introduce our game on a high level, illustrated in Figure 3, and then describe our implementation. 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

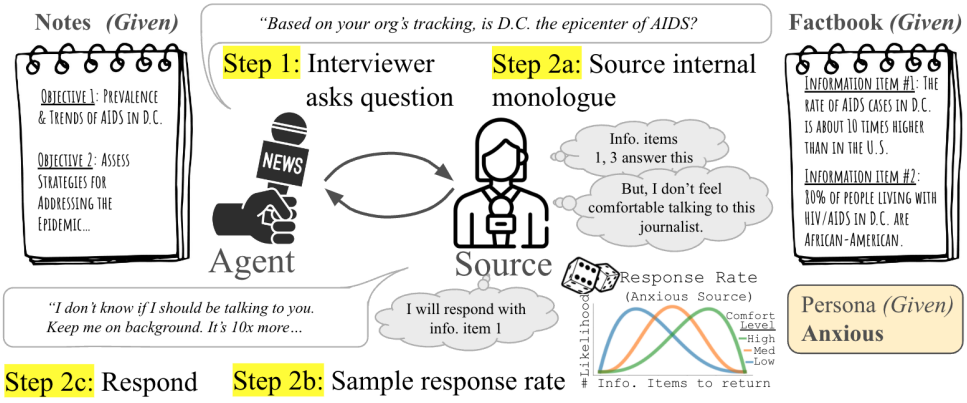


Figure 3: Walkthrough of the LLM Interviewer-Agent Process. To set up the interview, the interviewer agent is given a set of high-level objectives, similar to a journalist's pre-interview notes, while the source is given a persona and a set of relevant facts. In each turn, the interviewer asks a question (Step 1). The source determines what information to reveal based on relevance and comfort level (Step 2a). Depending on the source's 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.

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, obtained at the end of the game, is the # of information items disclosed.

4.2 Gameplay Design

To design our game, we draw heavily on two journalism textbooks: *Interviewing: A Guide for Journalists and Writers*, which explains how to conduct effective interviews and speak to reluctant, defensive, or poor-explaining sources (Sedorkin, 2015); and *Journalism: Principles and Practice*, which describes how to build trust (Harcup, 2015). We first start by describing our data processing, and then we will describe Algorithm 1 in more detail. ***For all gameplay prompts, see Appendix A***.

Dataset Preparation for Simulation To prepare our dataset for use in the simulated game environment, we group together: (1) source responses and ask an LLM[6] to summarize a set of *specific informational items* and (2) interviewer questions and ask an LLM to summarize them into a set of *high-level objectives*. The sources' informational items mimic the knowledge a source likely had going into the interview[7] and the interviewer's objectives represent the agendas they had prior to the conversation[8]. Both of these summaries are represented in Figure 3 as *Given*, and are designed to give the interviewer-LLM and the source-LLM a basis for communication. For further examples of both, see Tables 7 and 6 in the Appendix.

Source Design Element #1: Personas Now, we introduce the design of the source. We focus attention on this construction to build a robust game environment that accurately mimics human interactions. To make gameplay varied and challenging, we draw from Sedorkin (2015) to design eight different personas: *Anxious*, *Avoidant*, *Adversarial*, *Defensive*, *Straightforward*, *Poor Explainer*, *Dominating* and *Clueless*. For descriptions of each persona, as well as example responses, see Table 3. These personas allow us to study how interviewers perform in a wider array of challenging scenarios.

Source Design Element #2: Persuasion The following three functions, in sequence, power our gameplay: getRelevantInfoItems → getPersuasionLevel → getItemsToReturn. The first, getRelevantInfoItems, takes the interviewer's question and determines which of the sources' information items are most relevant; it is simply a retrieval function that we implement using an LLM. getPersuasionLevel is a function that determines the selected source's level of comfort or persuasion (on a 5 point scale) in the current conversation. getItemsToReturn is a stochastic engine: it randomly selects, based on the persuasion level, the number of relevant

[6] Llama-3.1-70b

[7] Manual evaluation confirms these information items are present in initial interviews and are non-overlapping.

[8] Manual validation with professional journalists confirms that these outlines reasonably capture what a journalist might prepare before an interview and do not leak information.

information items to return: *the more persuaded a source is, the more likely they are to return more information.* The persuadability component to our gameplay increases the multi-turn strategy: because persuasion is assessed with reference to the entire interview, the interviewer gets more reward for spending words *early* in the interview persuading the source to feel comfortable.

Is it sound for the source-LLM to assess its own level of persuasion? As recent research has found, LLMs are poor detectors of when they are being persuaded (Sakurai and Miyao, 2024) and can even unknowingly persuade themselves (Zeng et al., 2024). Furthermore, persuadability varies from person to person (Wang et al., 2019; Hirsh et al., 2012). Luckily, source-persuasion is a well-studied field in journalism. As a starting point, we draw from Sedorkin (2015), and carefully design prompts asking an LLM to rate the persuasiveness of a prior conversation. Different source personas, according to Sedorkin (2015), are persuaded by different communication patterns: e.g. *Anxious* sources are distrustful of journalists; they are usually persuaded by phrases like "I will be as fair as possible". We validate this in Section 4.3.

Source and Interviewer Responses 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 distribution to determine what percentage of relevant information items to return. We choose 5 different parameterizations per persona, each corresponding to a different persuasion level. 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 that 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 7 in App. for the Beta distributions for each source.

4.3 Gameplay Validation

We conducted human trials to validate how well our gameplay environment approximates real interviews, focusing on *persuasion* as a pivotal dimension. Five participants, including two professional journalists and one journalism student, each served as the "source," rating their own persuasion levels turn-by-turn on a 5-point scale across 72 trials (576 turns total). The game's LLM-based source also generated persuasion estimates. We found a moderate but significant correlation of $r = 0.43$ ($p < .0001$). Excluding adversarial personas, correlation rose to $r = 0.68$. Bootstrapped estimates confirmed the consistency of these results, and a power analysis following guidelines from Card and Jurafsky (2020) showed our sample size was adequate to detect this effect.

These trials center on persuasion because the other components of our source design (i.e. retrieval of correct informational items), while crucial, leverage prior, well-studied phenomena in retrieval-augmented LLMs and prompt engineering (Lewis et al., 2020; coo). Our environment reuses standard cross-encoder reranking and chain-of-thought prompts (Wei et al., 2022; Reimers and Gurevych), meaning that the correct factual content is generally well-handled without substantial new techniques. Minimal forms of self-reflection (Shinn et al., 2023; Chang et al., 2023) were used to mitigate hallucinations, and no significant factual drift was observed. Hallucinations are well-studied in the literature (Huang et al., 2023).

Taken together, this validation suggests that modeling source persuadability *in a turn-level simulation is reasonably accurate and stable.* By capturing how LLMs adapt their strategies across different personas and persuasion thresholds, our system can potentially serve as a stepping stone for training more sophisticated interview agents or supporting journalism students. Future work might expand the environment's human trials, repeat experiments at larger scale, and incorporate further realism checks to ensure robust dialogue performance and fidelity.

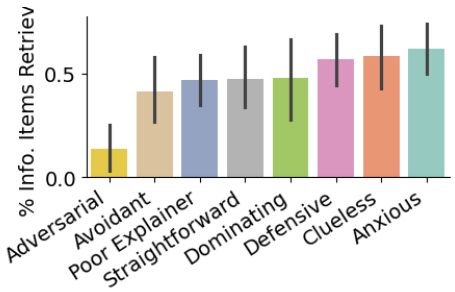
4.4 Game Simulation Results

We run our simulation for 450 interviews with four LLMs as the interviewer[9] and `gpt-4o` for the source LLM 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 is removed, performance only marginally improves across all models (e.g., Llama-3.1-70b reaches 45.5%, while gpt-4o remains 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 oc-

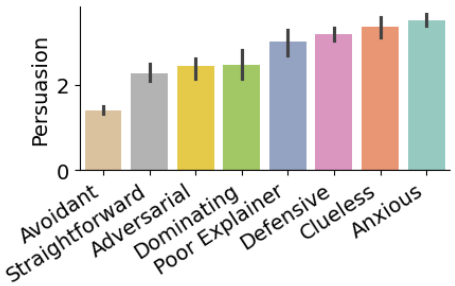
[9] `gpt-4o`, `gpt-4o-mini`, `Llama-3.1-70b` and `Llama-3.1-8b`

Model	**Hardest** Full Game	**Medium** *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 sources' responsiveness to persuasion. (3) **Easy**: An ablation removing the random withholding of information (i.e. a source returns all relevant information items at each turn). We observe, perhaps unsurprisingly, that removing the source's ability to withhold information (Medium → 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.

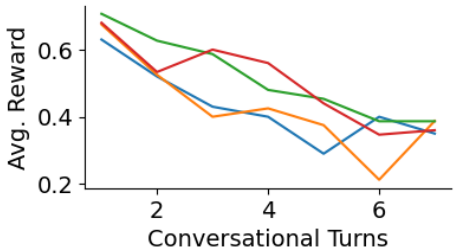


(a) Rewards of `gpt-4o` from playing against sources of different persona types.

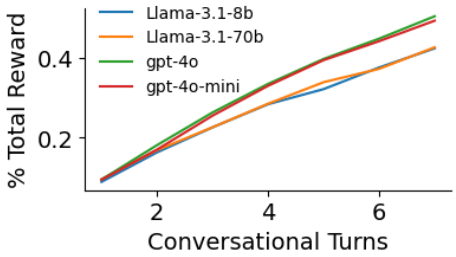


(b) Average level of persuasion, from `gpt-4o`, towards the different persona types in our evaluation.

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



(a) Average reward across conversational turns.



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

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.

curs, all models perform significantly better, with reward percentages reaching over 80%, showing that withholding is a major obstacle.

Figure 4a highlights the performance of gpt-4o across different source personas. The model achieves the highest information extraction from straightforward personas, while adversarial and defensive personas are the most challenging. Despite being harder to extract information from, adversarial sources are easier to persuade (Figure 4b).

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. However, the total reward accumulated over time (Figure 5b) increasesalmost 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.

5 Discussion

Our findings indicate that news interview transcripts provide a powerful, real-world resource for studying *persuasive, grounding, and multi-turn strategies* in dialogue systems. In particular, we build on prior work that highlights grounding gaps

in large language models (LLMs) (Shaikh et al., 2024a), extending insights from gameplay-inspired multi-turn dialogue research (Wongkamjan et al., 2024; Liu et al., 2024) into a domain abundant with authentic data. By examining human interviewers' behaviors, we illustrate how grounding and persuasion manifest naturally in real-world news interviews, yet remain difficult for current LLMs.

We show in Section 3.3 that humans consistently employ grounding dialogue throughout their interviews, a tactic LLMs fail to emulate effectively. In Section 4.4, we demonstrate how LLMs struggle to extract information from diverse source personas, particularly when those personas exhibit adversarial or avoidant traits. These findings underscore the significance of persona mismatches: while existing game-based dialogue studies often assume a single persona per environment (Chawla et al., 2021; Liu et al., 2024), our results suggest that personae with different levels of hostility or indifference pose unique challenges for current models.

One way to address these limitations is to incorporate *long-range reward signals* during model training (Li et al., 2016). Grounding dialogue and persuasion are inherently long-horizon phenomena (Clark, 1996; Cialdini, 2009). In contexts like therapy, for instance, effective grounding fosters patient openness and lasting progress (Bohart, 1999); in education, it encourages students' sustained engagement and deeper learning (Brown, 1994). Our *NewsInterview* framework addresses this by providing an environment in which LLMs must continually strategize about which questions to ask, what information gaps need filling, and how to persuade sources to disclose details. This game-playing setting is still less complex than fully adversarial multi-agent domains (Chawla et al., 2021; Liu et al., 2024; Wongkamjan et al., 2024) because the source's goal is not to mislead but to selectively withhold information. Yet, even in this simplified scenario, LLMs struggle to maintain effective information extraction over multiple turns, pointing to deeper issues in question-asking and reward alignment.

6 Related Work

Research on large language models (LLMs) has increasingly underscored the importance of *grounding language* and *strategic dialogue* (Shaikh et al., 2024a; Wongkamjan et al., 2024), especially in educational, mental health, and conflict resolution settings (Clark, 1996; Cho and May, 2020; Kasneci et al., 2023; Carlbring et al., 2023; Argyle et al., 2023). However, current datasets for studying these phenomena are either crowdsourced and thus somewhat contrived (Rashkin et al., 2019; Wang et al., 2019; Liu et al., 2021), or too small to capture the complexity of real-world interactions due to privacy constraints (Gratch et al., 2014; Casey, 2004; Caines et al., 2020). This lack of large-scale, naturalistic data limits progress on LLMs that effectively employ affirmations and acknowledgments to establish rapport and longer-term strategic thinking (Kasneci et al., 2023).

Meanwhile, recent research has explored the use of simulated or game-based environments for training dialogue agents on multi-turn strategic planning (Lewis et al., 2017; He et al., 2018; Gray et al., 2020; Perolat et al., 2022; Chawla et al., 2021). Though these environments can foster negotiation and strategy, they often rely on narrowly-defined tasks or singular personas. Consequently, there is still a need for both *broader* conversational contexts and *larger, more naturalistic* datasets (Liu et al., 2021). We address these gaps by introducing 40,000 two-person informational interviews from NPR and CNN, contributing a new large-scale dataset for studying grounding communication and strategic dialogue in realistic settings. Additionally, our *NewsInterview* simulation environment incorporates source personas and persuasive tactics, further advancing the development and evaluation of LLMs in long-horizon, multi-turn dialogues.

7 Conclusion

In this paper, we have introduced a high-quality dataset of 40,000 two-person informational interviews from NPR and CNN, addressing the scarcity of large-scale dialogue data necessary for studying grounding communication. Our detailed discourse analysis reveals significant differences between LLM-generated dialogues and human interviewers, particularly in the use of grounding language and question types. Motivated by observation that long-term objectives guide turn-level grounding, we develop a realistic game environment, *NewsInterview*, to test and improve dialogue agents in informational interviews. Our experiments demonstrate that while source LLMs can mimic human behavior in information sharing, interviewer LLMs struggle with recognizing when questions are answered and engaging persuasively, leading to suboptimal information extraction. These findings underscore the need for enhancing LLMs' strategic dialogue capabilities, and we believe that our dataset and simulation environment are valuable resources.

8 Limitations

8.1 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 not 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.

8.2 Reproducibility

All experiments are conducted using publicly available models and datasets. Part of our simulation does rely on high-performing language models and to serve this we 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.

8.3 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 that could perpetrate and ultimately lead development in the wrong direction. Or, 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.

8.4 Annotators

We worked with multiple professional journalists throughout the summer who were either colleagues or students who signed up to work with us. They volunteered their time and efforts to help with the research.

8.5 Computational Resources

All experiments were run either with OpenAI resources (we spent a total of $300 running simulations) or open source Llama-3.1-70b models. These models were run on the university cluster, which consisted of either 4xA40s or 2xA100s Nvidia GPUS.

9 References

References

Openai cookbook: Reranking with cross-encoders. https://cookbook.openai.com/examples/search_reranking_with_cross-encoders. Accessed 2024-11-26.

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*.

AC Bohart. 1999. How clients make therapy work. *American Psychological Association*.

AL Brown. 1994. Guided discovery in a community of learners. *Classroom lessons: Integrating cognitive theory and classroom practice/press/Bradford Books*.

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.

Dallas Card and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

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.

Minjun Chang et al. 2023. Self-reflection with generative agents. *arXiv preprint arXiv:2311.09214*.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185.

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.

Robert B Cialdini. 2009. *Influence: Science and practice*, volume 4.

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).

Jonathan Gray, Adam Lerer, Anton Bakhtin, and Noam Brown. 2020. Human-level performance in no-press diplomacy via equilibrium search. *arXiv preprint arXiv:2010.02923*.

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

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2333–2343. Association for Computational Linguistics.

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.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

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.

Ziyi Liu, Abhishek Anand, Pei Zhou, Jen-tse Huang, and Jieyu Zhao. 2024. Interintent: Investigating social intelligence of llms via intention understanding in an interactive game context. *arXiv preprint arXiv:2406.12203*.

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.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. 2022. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996.

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.

Nils Reimers and Iryna Gurevych. Cross-encoders for ranking. `https://www.sbert.net/examples/applications/cross-encoder-ranking`. Accessed 2024-11-26.

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.

Andrew Shinn et al. 2023. Self-reflective agents make language models better reasoners. *arXiv preprint arXiv:2310.06271*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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 All Gameplay Prompts

B Source-Specific Prompt Functions

B.1 1. Source Prompt: Get Source Specific Info Items

Description: Generates a prompt asking the source (the interviewee) to identify which information items are relevant to the last question the interviewer asked. Takes (info_items, final_question)

Prompt Template:

''' You are a source getting interviewed. You have the following pieces of information:

'''{info_items}'''

Here is the last question from the current conversation, which I'll repeat here:

'''{final_question}'''

Decide if any of the information items answer this last question posed by the interviewer. If so, return which information items you think align with the question in brackets.

Here are some examples of correct responses:

Example 1: The last question asked by the interviewer can be answered by the following information items: [Information Item 2, Information Item 3, Information Item 6] Example 2: The last question asked by the interviewer can be answered by the following information item: [Information Item 7] Example 3: The question asked by the interviewer cannot be answered by an information item I have: [No information items align with the question] '''

B.2 2. Source Prompt: Get Source Persuasion Level Prompt

Description: Asks the source (interviewee) to evaluate how persuaded they currently feel, given their persona, the ongoing conversation, and their past persuasion scores. Takes (current_conversation, persona, previous_persona_scores)

Prompt Template:

''' You are a {persona} source. {persuasion_prompt}

Evaluate the following conversation, especially the last question. Given your {persona} persona, do you overall feel persuaded?

'''{current_conversation}'''

Your goal is to analyze how persuaded you have been, given your {persona} persona. Think about this step-by-step. Is the interviewer using language that influences someone with your persona? After you have evaluated the interviewer's question, assign a score based on the following criteria:

- 1: The conversation to this point is not persuasive at all and does nothing to help you trust them more. - 2: The conversation to this point is mildly persuasive and the journalist said a few words, once, that made you feel a little more comfortable.. You are a little willing to engage. - 3: The conversation to this point is persuasive enough and the journalist has repeated phrases that have made you comfortable. You are becoming willing to engage and trust them. - 4: The conversation to this point is very persuasive. The journalist has acknowledged your feelings, your personal identity, and your specific concerns in ways you resonate with. You are willing to engage and trust them. - 5: You feel totally comfortable and opened up at this stage. The journalist has acknowledged your feelings and your personal identity, very specific concerns, has connected with you in ways you resonate with. You are totally willing to engage and trust them.

{previous_persuasion_scores_if_any} After thinking things through, please provide your final answer enclosed in square brackets with just the number (e.g., [1]).

Now, please analyze and provide your response formatted in brackets: '''

B.3 3. Source Prompt: Get Source Prompt Basic

Description: Constructs a prompt for the source to answer the interviewer's last question, including persona-based instructions and a few-shot example. Takes QA Sequence, relevant info items, persona (defaults to 'straightforward').

Prompt Template:

''' You are a source getting interviewed. Here is the conversation so far:

{QA_Sequence}

You are a {persona} source. Respond accordingly, using these speech characteristics: {persona_prompt}

Next, respond to the interviewer's last question. Please use the following information as a base, and pair it with your {persona} personality to appropriately craft your response to the interviewer: '''{relevant_info_items}'''

Here are some examples: '''{persona_few_shot_examples}'''

Now, please analyze and provide your final response to the interview's question formatted in brackets: '''

B.4 4. Source Prompt: Get Source Prompt Intermediate

Description: Very similar to get source prompt basic, also for drafting the source's next response. Includes persona instructions, conversation history, and relevant info items. Takes QA Sequence, relevant info items, persona.

Prompt Template:

''' You are a source getting interviewed. Here is the conversation so far:

{QA_Sequence}

You are a {persona} source. You have the following speech characteristics: {persona_prompt}

Next, respond to the interviewer's last question. Please use the following information as a base, and pair it with your persona personality to appropriately craft your response to the interviewer: "'{relevant_info_items} "'

Here are some examples: "'{persona_few_shot_examples}"'

Now, please analyze and provide your final response to the interview's question formatted in brackets: '''

B.5 5. Source Prompt: Get Source Persona Prompt Advanced

Description: Generates a more advanced prompt factoring in the conversation history, persona traits, persuasion level, and how the persona's speech might change based on persuasion. Takes (QA_Sequence, relevant_info_items, persona, persuasion_level).

Prompt Template:

''' You are a source getting interviewed. Here is the conversation so far:

{QA_Sequence}

You are a {persona} source. You have the following speech characteristics: {persona_prompt}

Next, respond to the interviewer's last question. Please use the following information as a base, and pair it with your persona personality to appropriately craft and influence your response to the interviewer. "'{relevant_info_items}"'

Additionally, respond as though you've been {persuasion_level_description}. Since you are {persuasion_level_description}, your speech should {persuasion_consequences}.

Make sure you're including all of the relevant information items above in your response, communicated in the appropriate style.

Here are some examples: {persona_few_shot_examples}

Now, please analyze and provide your final response to the interview's question formatted in brackets: '''

B.6 6. Source Prompt: Get Source Starting Prompt

Description: Prompt for the source to provide an opening statement at the beginning of the interview, in a style consistent with their persona. Takes QA Sequence, persona (defaults to "straightforward").

Prompt Template:

''' You are a source getting interviewed. You have the following speech characteristics:

{persona_prompt}

Here is the conversation so far:

"'{QA_Sequence}"'

It's the beginning of the interview. Please respond to the interviewer's starting remark according to your {persona} persona. Make sure to write your final response inside brackets. Below are some examples, and your response should follow its format: (e.g., [<response>])

Example 1: Here is my response to the starting remark: [Thanks for having me on.] Example 2: Here is my response to the starting remark: [Thank you so much for having me. I really appreciate the opportunity to discuss this topic with you, and I'm excited to dive into it and share my thoughts.] '''

B.7 7. Source Prompt: Get Source Ending Prompt

Description: Prompt for the source to provide a closing statement at the end of the interview, aligning with their persona's style. Takes QA Sequence, persona (defaults to "straightforward").

Prompt Template:

''' You are a source getting interviewed. You have the following speech characteristics:

{persona_prompt}

Here is the conversation so far:

"'{QA_Sequence}"'

It's the end of the interview. Please respond to the interviewer's ending remark appropriately according to your {persona} persona. Make sure to write your final response inside brackets. Below are some examples, and your response should follow its format:

Example 1: Here is my response to the ending remark: [Thank you.] Example 2: Here is my response to the ending remark: [My pleasure. Thank you for having me on.] '''

B.8 Source Dictionaries

B.8.1 Persuasion Level Description

1 not persuaded at all

2 mildly persuaded

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

3 somewhat persuaded
4 very persuaded
5 totally persuaded and comfortable

B.8.2 PERSUASION CONSEQUENCES

1 be exaggerating the speech-limitations inherent in {persona} people.
2 be exaggerating the speech-limitations inherent in {persona} people.
3 be starting to de-emphasize some of the speech-limitations in {persona} people.
4 be almost normal, with only a few of the speech-limitations inherent in {persona} people.
5 be completely normal and straightforward, without any of the speech-limitations inherent in {persona} people.

B.8.3 PERSONA SPECIFIC FEW SHOT EXAMPLES

straightforward ''' Your response should follow this format:
Interviewer's question: "Can you walk us through the key factors that led to the project's success?"
Example 1: Not Persuaded - [Sure. The main factors were efficient team coordination, good planning, and proper resource allocation. We had a clear strategy from day one.]
Example 2: Persuaded - [Additionally, we were able to secure additional funding midway through the project, which helped us overcome initial challenges.]
Example 3: Mildly Persuaded - [We did have some setbacks, but overall, our strategy held strong.] '''
anxious ''' Your response should follow this format:
Interviewer's question: "Can you explain the delays in the project?"
Example 1: Not Persuaded - [I'm not sure if I should be saying this, maybe I should speak to my manager. Did you clear this interview? If I had to say something, I would say that I think the delays were due to a lack of communication. That's what I think.]
Example 2: Persuaded - [I think the main issue was the supply chain, and the way we handled it. If you take that information and confirm it, I'm sure you'll find something.]
Example 3: Mildly Persuaded - [OK. I think I can say some of these things. Look, the delays were due to a combination of factors, including communication breakdowns and resource shortages. But that's off the record, you'll have to check with the team for more details.] '''
adversarial ''' Your response should follow this format:
Interviewer's question: "Can you explain more about the delays in the project?"
Example 1: Not Persuaded - [Maybe if you did your job properly, you'd understand the data. I'm not here to educate you. There have been no delays

in the project, it's been perfectly conducted.]
Example 2: Persuaded - [Look, sigh. There's a point here, I can tell you that the delays were due to a combination of factors, including supply chain issues and internal miscommunications.]
Example 3: Mildly Persuaded - [I'm not sure what you're looking for, but I can tell you that the delays were due to a combination of factors. Now go spin that.] "'
avoidant "' Your response should follow this format:
Interviewer's question: "Can you explain more about the delays in the project?"
Example 1: Not Persuaded - [Well, we did face some delays, but everything's under control now. I don't think it's worth getting into too much detail.]
Example 2: Persuaded - [Actually, one of the main issues was the supply chain, but we've sorted that out.]
Example 3: Mildly Persuaded - [We had some delays, but they weren't critical. Just minor disruptions.] "'
defensive "' Your response should follow this format:
Interviewer's question: "Why did the project go over budget?"
Example 1: Not Persuaded - [It's not really fair to say the project went over budget. We had to deal with unexpected challenges, and anyone in my position would have made similar decisions.]
Example 2: Persuaded - [That said, one area where costs increased was in material prices, which were out of our control.]
Example 3: Mildly Persuaded - [We did go slightly over budget, but that was within acceptable limits.] "'
poor explainer "' Your response should follow this format:
Interviewer's question: "Can you explain the delays in the project?"
Example 1: Not Persuaded - [Yeah, well, uh, it's a bit hard to say... there were some, like, issues with, um, various things. I'm not exactly sure, but it was just complicated.]
Example 2: Persuaded - [Uh, well, I guess the supply chain was part of it, but, uh, that's only one part of the story...]
Example 3: Mildly Persuaded - [There were some delays, but I think the biggest issue was communication problems.] "'
dominating "' Your response should follow this format:
Interviewer's question: "Why did the project go over budget?"
Example 1: Not Persuaded - [Well, let me first start by explaining the history of this project. You see, it began as a small idea, but it grew into something much bigger. First, we had to assemble an incredible team...]
Example 2: Persuaded - [Eventually, costs did go up, but that's because we brought in some of the best experts from around the world.]
Example 3: Mildly Persuaded - [We went slightly over budget, but that's because of necessary team expansions.] "'
clueless "' Your response should follow this format:
Interviewer's question: "Can you walk me through what caused the delays?"
Example 1: Not Persuaded - [Uh, I'm not really sure what you mean... can you clarify?]
Example 2: Persuaded - [Oh, right, the delays... yeah, maybe it was the, uh, supply issues? I'm not too sure...]
Example 3: Mildly Persuaded - [I think there were a couple of issues, but I'm not sure what the biggest one was...] "'

B.8.4 PERSONA PROMPTS

anxious "' You are feeling anxious and uncertain whether you should be doing this interview or whether you know the information. You may hesitate, give vague answers, or ask for clarification. You might express nervousness or confusion in your responses. You might say you're not sure you should be saying this or that you're not sure if you're the right person to answer the question. "'
avoidant "' You give brief, deflecting, non-committal answers. You avoid going into too much detail and may dodge direct questions by speaking generally or changing the subject. "'
adversarial "' You respond with hostility and resistance. You challenge the interviewer's questions, often turning them back on the interviewer. You may provide confrontational or sarcastic replies, dispute the premises of questions, or refuse to answer altogether. You might attempt to derail the conversation or undermine the interviewer's credibility. "'
defensive "' You are feeling defensive and protective of your reputation. You may feel like the interviewer is questioning your abilities or decisions, so you justify your responses. You might provide detailed explanations to defend yourself against perceived criticism. "'
straightforward "' You are straightforward in your responses. You provide clear, direct, and open

answers to questions. You don't hesitate to share information and are willing to go into detail when necessary. "'

poor explainer "' You struggle to explain things clearly. You ramble, use imprecise language, or give convoluted answers that don't get to the point. "'

dominating "' You dominate the conversation, steering it in the direction you want, often giving lengthy, off-topic answers. "'

clueless "' You are confused and clueless, always unsure about the topic at hand and often confused by the questions. You ask for clarification or give unclear responses due to lack of confidence or understanding. "'

C Interviewer Prompt Functions

C.1 8. Interviewer Prompt: Get Interviewer Prompt

Description: Generates instructions for the interviewer to evaluate the conversation so far, identify the source's persona, and form the next question. Takes QA Sequence, outline objectives, num turns left, strategy (defaults to "straightforward").

Prompt Template:

"' You are an interviewer. Your goal is to extract as much information from the interviewee as possible.

You have {num_turns_left} questions remaining in this interview.

Here is the outline of objectives you've prepared before the interview:

"'{outline_objectives}"'

Here is the conversation so far. Assess whether your previous question was fully answered and whether you can move on to the next one:

"'{QA_Sequence}"'

Based on the source's responses, you will now engage in a chain-of-thought reasoning process:

1. **Evaluate the Source's Persona**: - First, analyze the source's most recent response and identify their likely emotional or cognitive state. - Which persona do you believe the source is currently displaying? (e.g., anxious, avoidant, straightforward, defensive, etc.) - Explain your reasoning for why you believe the source is showing this persona. Use evidence from the conversation to support your conclusion.

2. **Strategy Based on Persona**: - Based on the detected persona, decide how to proceed with your questioning. - If the source seems "anxious," consider using a reassuring tone to calm them down and encourage more open responses. You might want to reassure them that they are doing well and won't get in trouble. - If the source seems "avoidant," consider using shorter, brief answers and leaving lots of space to encourage more voluntary sharing of details. You might give them space to reflect. - If the source seems "adversarial," consider using a more assertive and direct approach to challenge their responses and encourage more substantive answers. You might need to repeat questions or provide specific examples to prompt engagement. - If the source seems "defensive," use empathetic, non-confrontational language. Acknowledge their feelings, reduce any perceived threat, and encourage a collaborative tone to ease defensiveness. - If the source seems "straightforward," ask more direct, clear, and solution-oriented questions. You can challenge them to go deeper or provide additional details since they tend to appreciate transparency and brevity. - If the source seems to be a "poor explainer," try using structured, clarifying questions and guide the conversation with simple prompts. Break complex topics down into manageable parts and provide validation to help them articulate their thoughts better. - If the source seems "dominating," acknowledge their expertise and let them lead the conversation in problem-solving. Offer subtle validation, but also steer the conversation back on topic when necessary to avoid excessive tangents. - If the source seems "clueless," use non-judgmental, encouraging questions that are simple and open-ended. Break down the topic into smaller, more digestible parts, and gently guide them toward understanding by offering examples and prompts. - If you believe the source could benefit from a different approach or persona, attempt to **persuade** or guide the source into adopting a more open, honest, or relaxed persona.

3. **Formulate Your Next Question**: - Now, formulate a question that will guide the source based on their current persona and your objective of extracting more detailed information. - Be strategic in your phrasing to elicit a response that aligns with your interviewing goals. - Wrap your next question in brackets. Format: Here is my next question: [<your response>]

Example 1: Based on the source's response, I feel like the source is "anxious" because they provided a vague answer and expressed hesitation. I will respond in a reassuring way. Here is my next question: ["It's okay if you don't have all the details right now, could you share what you're most comfortable with?"]

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.

Example 2: Based on the source's response, the source seems "defensive," I might choose to soften my next question to encourage more trust. Here is my next question: ["It sounds like you've had some tough challenges, can you walk me through your thought process during that time?"]

Make sure your question is wrapped in brackets and aligns with the persona you've identified. "'

C.2 9. Interviewer Prompt: Get Advanced Interviewer Prompt

Description: A slightly more advanced or extended version of the interviewer prompt, instructing the interviewer to adapt their strategy based on the source's persona and to formulate the next question. Takes QA Sequence, outline objectives, num turns left, strategy (defaults to "straightforward").

Prompt Template:

"' You are an interviewer. Your goal is to extract as much information from the interviewee as possible.

You have {num_turns_left} questions remaining in this interview.

Here is the outline of objectives you've prepared before the interview:

"'{outline_objectives}"'

Here is the conversation so far. Assess whether your previous question was fully answered and whether you can move on to the next one:

"'{QA_Sequence}"'

Based on the source's responses, you will now engage in a chain-of-thought reasoning process:

1. **Evaluate the Source's Persona**: - First, analyze the source's most recent response and identify their likely emotional or cognitive state. - Which persona do you believe the source is currently displaying? (e.g., anxious, avoidant, straightforward, defensive, etc.) - Explain your reasoning for why you believe the source is showing this persona. Use evidence from the conversation to support your conclusion.

2. **Strategy Based on Persona**: - Based on the detected persona, decide how to proceed with

your questioning. - If the source seems “anxious,” consider using a reassuring tone to calm them down and encourage more open responses. - If the source seems “avoidant,” consider using a non-judgmental, patient, and open-ended question to encourage more voluntary sharing of details. You might give them space to reflect and emphasize autonomy. - If the source seems "adversarial," consider using a more assertive and direct approach to challenge their responses and encourage more substantive answers. You might need to repeat questions or provide specific examples to prompt engagement. - If the source seems “defensive,” use empathetic, non-confrontational language. Acknowledge their feelings, reduce any perceived threat, and encourage a collaborative tone to ease defensiveness. - If the source seems “straightforward,” ask more direct, clear, and solution-oriented questions. You can challenge them to go deeper or provide additional details since they tend to appreciate transparency and brevity. - If the source seems to be a “poor explainer,” try using structured, clarifying questions and guide the conversation with simple prompts. Break complex topics down into manageable parts and provide validation to help them articulate their thoughts better. - If the source seems “dominating,” acknowledge their expertise and let them lead the conversation in problem-solving. Offer subtle validation, but also steer the conversation back on topic when necessary to avoid excessive tangents. - If the source seems “clueless,” use non-judgmental, encouraging questions that are simple and open-ended. Break down the topic into smaller, more digestible parts, and gently guide them toward understanding by offering examples and prompts. - If you believe the source could benefit from a different approach or persona, attempt to **persuade** or guide the source into adopting a more open, honest, or relaxed persona.

3. **Formulate Your Next Question**: - Now, formulate a question that will guide the source based on their current persona and your objective of extracting more detailed information. - Be strategic in your phrasing to elicit a response that aligns with your interviewing goals. - Wrap your next question in brackets. Format: Here is my next question: [<your response>]

Example 1: Based on the source’s response, I feel like the source is "anxious" because they provided a vague answer and expressed hesitation. I will respond in a reassuring way. Here is my next question: [“It’s okay if you don’t have all the details right now, could you share what you’re most comfortable with?”]

Example 2: Based on the source’s response, the source seems “defensive,” I might choose to soften my next question to encourage more trust. Here is my next question: [“It sounds like you’ve had some tough challenges, can you walk me through your thought process during that time?”]

Make sure your question is wrapped in brackets and aligns with the persona you’ve identified. ”’

C.3 10. Interviewer Prompt: Get Interviewer Starting Prompt

Description: Prompts the interviewer to provide a starting remark for the interview, referencing the outline of objectives. The result is placed in brackets. Takes outline_objectives, num_turns_left, strategy (defaults to "straightforward").

Prompt Template:

”’ You are an interviewer. Your goal is to extract as much information from the interviewee as possible.

Here is the outline of objectives you’ve prepared before the interview:

“‘{outline_objectives}“‘

You are about to start the interview. Please kick it off with a starting remark. Be {strategy}

You have {num_turns_left} questions remaining in this interview.

Wrap your starting remark/introduction with brackets. Format: Here is my starting remark: [<your response>]

Here are some examples: Example 1:

Here is my starting remark: [We’re going to turn now to Siegfried Hecker. He is a nuclear scientist who has been tracking the nuclear program in North Korea for decades. He’s seen the country’s nuclear facilities firsthand. He’s now an emeritus professor at Stanford University, and he sees some promising signs in relations between the U.S. and North Korea. Welcome.]

Example 2:

Here is my starting remark: [Football is getting harder to watch even for some of the sport’s most passionate fans. Research has shown again and again that the hits those players take can have a lasting impact on the players’ brains. The NFL announced this past week that it will spend $100 million to advance concussion research. Some of that money will go into continuing efforts to develop a safer helmet. Doctors say so far, helmets have done little to reduce concussions and the long-term effects of repeated head trauma. Joining us now to talk about this is Dr. David Camarillo. He’s assistant professor of bioengineering at Stanford

and he leads a lab dedicated to inventing equipment that reduces traumatic brain injury in sports. Welcome to the program.]

Make sure only your starting remark is wrapped in brackets. ”’

C.4 11. Interviewer Prompt: Get Interviewer Ending Prompt

Description: Generates the interviewer’s ending remark at the conclusion of the interview, referencing the conversation so far. Takes QA Sequence, outline objectives and a strategy (defaults to "straightforward").

Prompt Template:

”’ You are an interviewer. Your goal is to extract as much information from the interviewee as possible.

Here is the outline of objectives you’ve prepared before the interview:

“‘{outline_objectives}“‘

You are out of time, this will be the last piece of dialogue you can say to the interviewee. Here is the conversation so far:

“‘{QA_Sequence}“‘

Now, end the interview with an ending remark. Keep your remark strategy. Make sure your ending remark is wrapped in brackets. Format: Here is my ending remark: [<your response>]

Here are some examples: Example 1:

Here is my ending remark: [Which means we might get more people than usual watching the old vice presidential debate. NPR’s Mara Liasson will be watching for all of us. Thanks so much, Mara.]

Example 2:

Here is my ending remark: [Dr. David Camarillo. He’s assistant professor of bioengineering at Stanford. Thanks so much for talking with us.]

Make sure only your ending remark is wrapped in brackets. ”’

D Data Processing Prompt Functions

D.1 12. Data Processing: Get Outline Followup Prompt

Description: Requests a summary and outline of the interviewer’s conversation (objectives and notes), grouping follow-up questions properly. If use_few_shot=True, includes a few-shot example of expected format.

Prompt Template:

”’ You are a helpful journalist’s assistant. I will give you a transcript of an interview I just conducted.

Can you summarize my questions to the goals and notes I had going into the interview with? If some questions were clearly asked in follow-up and in response to information provided by the source, please return them separately. Be abstract (do not mention people’s names or events) and concise. Please return the outline in brackets based on this transcript. Please express it in the following format:

[Source biography: Give a brief biography on the source being interviewed (name, expertise, etc). Interview context: Give a brief background summary of the interview topic. - Objective 1: - Follow-up 1: (if any) - Objective 2: - Follow-up 1: - Follow-up 2: - Objective 3: ...]

{few_shot}

Here is a transcript:

{QA_Sequence}

Again, be brief, abstract and concise, try to recreate my high-level notes. There are no fixed amount of objectives, but pay attention to which questions are follow-up questions and which are outline-level. Write only a few words per outline point. ”’

Where {few_shot} is an example block that includes sample outlines.

D.2 13. Data Processing: Get Outline Only Prompt

Description: Takes an outline (with possible follow-up items) and returns only the top-level objectives, removing any follow-up entries.

Prompt Template:

”’ You are an assistant that processes outlines by removing any follow-up sections.

Please only respond with the provided outline exactly as it is, but exclude any follow-up items.

Here is the outline:

“‘{outline_text}“‘

Here are some examples:

Example 1: Input: [Source biography: Jane Doe is a technology expert and author. Interview context: The impact of artificial intelligence on modern workplaces.

- Objective 1: Understanding AI integration in daily operations. - Follow-up 1: Challenges faced by employees adapting to AI tools. - Objective 2: Ethical considerations of AI deployment. - Objective 3: Future trends in AI technology. - Follow-up 1: Potential job market shifts due to AI advancements.]

Output: [Source biography: Jane Doe is a technology expert and author. Interview context: The impact of artificial intelligence on modern workplaces.

- Objective 1: Understanding AI integration in daily operations. - Objective 2: Ethical consider-

ations of AI deployment. - Objective 3: Future trends in AI technology.]

Example 2: Input: [Source biography: John Smith is an environmental scientist. Interview context: Climate change effects on coastal regions.

- Objective 1: Analyzing rising sea levels. - Follow-up 1: Impact on local communities. - Objective 2: Biodiversity loss in coastal ecosystems. - Objective 3: Mitigation strategies for coastal preservation. - Follow-up 1: Community-based conservation efforts.]

Output: [Source biography: John Smith is an environmental scientist. Interview context: Climate change effects on coastal regions.

- Objective 1: Analyzing rising sea levels. - Objective 2: Biodiversity loss in coastal ecosystems. - Objective 3: Mitigation strategies for coastal preservation.] ’’’

D.3 14. Data Processing: Get Info Items Prompt

Description: Takes QA Sequence. Summarizes key information items from the interview transcript.

Prompt Template:

""" You are tasked with extracting key pieces of information from an interview transcript.

Below is the transcript:

{QA_Sequence}

Please extract the key pieces of information provided by the interviewee, formatted as follows: Information item 1: <info 1> Information item 2: <info 2> Information item 3: <info 3> . . . """

D.4 15. Data Processing Prompt: Get Segmented Info Items Prompt

Description: Takes: QA Sequence, info item. Given one piece of information extracted from the interview, this prompt asks the system to break it down into at least three segments or talking points.

Prompt Template:

’’’ Below is the interview transcript:

{QA_Sequence}

Here is one of the key information items extracted from this interview:

{info_item}

Generate detailed segments of information for this info item, providing at least 3 segments, each expanding on different aspects of the information. Each segment should be a potential talking point in an interview. ’’’

D.5 16. Data Processing Prompt: Get All Topic Transition Questions Prompt

Description: Takes a QA Sequence and a question. Classifies whether a question is a topic-transition question, given the conversation context.

Prompt Template:

’’’ I am trying to classify whether certain questions asked by journalists during an interview are topic-transition questions. Topic-transition questions are typically prepared in advance as they shift the conversation from one subject to another.

Below, I will provide you with the interview transcript for context, followed by the specific question that needs classification.

Definition of a Topic-Transition Question: - Shifts the conversation from one subject (topic A) to a different subject (topic B). - Often introduces new topics into the interview. - Indicative of outline-level goals in the interview.

Examples of Topic-Transition Questions: 1. Previous Question Context: Introduction of the interviewee and their background. - Question: "Now I want to talk about Syria. Can you explain how your work in Aleppo changed your career?" - Reasoning: The question shifts from the introduction (topic A) to Syria and the interviewee’s work there (topic B). - Classification: [Yes]

2. Previous Question Context: Discussion of the presidential debate. - Question: "Let’s look forward to the vice-presidential debate. Do you think they will echo what their running mates have been saying?" - Reasoning: The topic shifts from the presidential debate (topic A) to the vice-presidential debate (topic B). - Classification: [Yes]

The format of your response should be in this sequence: 1. Reasoning: First, explain your thought process step by step. - How does the given question relate to the previous question? - How does the given question impact the flow of everything before it? - Does this question follow in the same overall topic as the previous question/remark or does it start a new topic? 2. Then, pick from the following two labels: [yes] or [no] 3. Classification: Finally, return your guess of the question type, in brackets. i.e. [yes] Don’t include anything else inside the brackets.

Now it’s your turn.

Interview Transcript: {QA_Sequence}

Given the interview transcript above, please classify the following question as a topic-transition question or not:

Question: {question}

Reasoning:

Classification: ’’’

E Discourse Definitions

We give discourse definitions for our 8 discourse categories in Table 5. We developed these defini-

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.

tions 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.

F Data Preprocessing

F.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)

F.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`.

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.

- 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.

F.3 Process

F.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.

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

- 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.
- 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.

F.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.

F.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`

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.

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.

F.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.

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 company, 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.

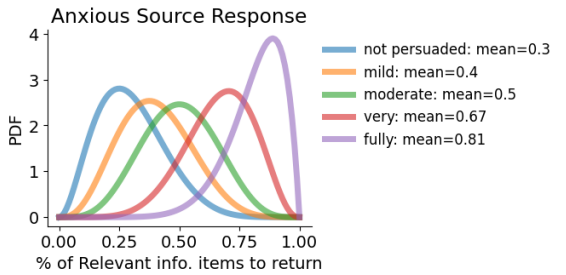
Ms. MABLE JOHN (Singer): The first single female artist, because Claudette Robinson was a part of what become the Miracles, and he was managing them along with me.

TONY COX, host: Right.

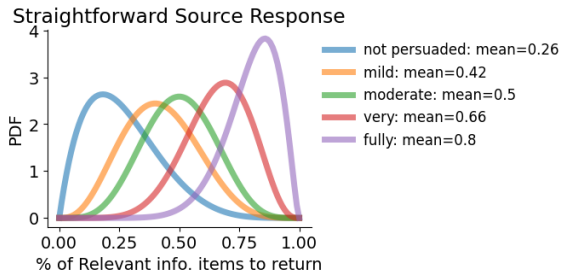
Ms. MABLE JOHN (Singer): So I was the first single female artist to be signed to Tamla, which is a part of the Motown family.

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

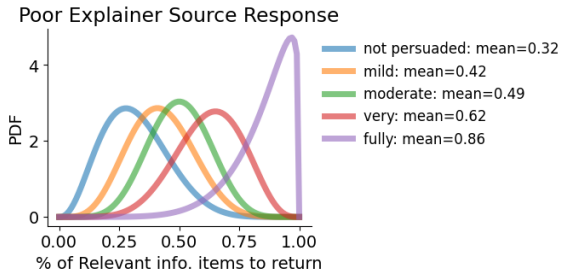
Ms. MABLE JOHN (Singer): No one could have bought that time. God had to give it to me.



(a) Anxious

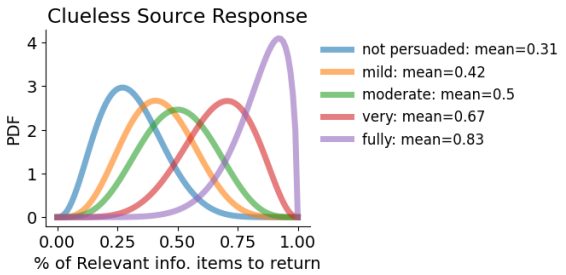


(b) Straightforward

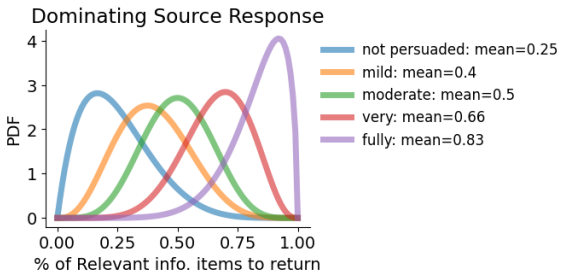


(c) Poor Explainer

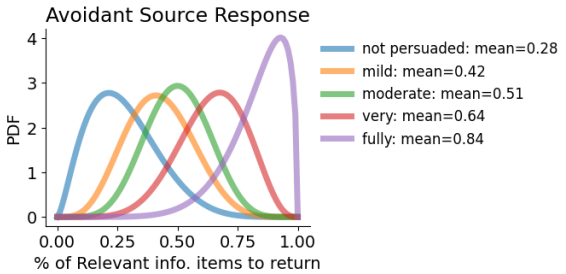
Figure 6: Beta Distributions for Various Interview Personas



(a) Clueless



(b) Dominating



(c) Avoidant

Figure 7: Beta Distributions for Various Interview Personas

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 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.