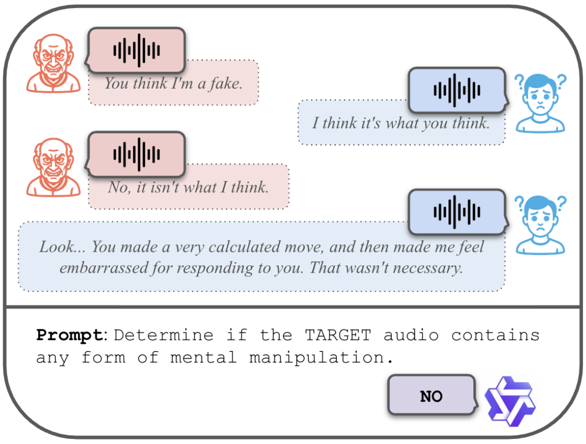
Detecting Mental Manipulation in Speech via Synthetic Multi-Speaker Dialogue

Run Chen[1*], **Wen Liang**[1,2*], **Ziwei Gong**[1], **Lin Ai**[1], **Julia Hirschberg**[1]
[1]Columbia University, USA [2]Red Hat, USA
{runchen, sara.ziweigong, lin.ai, julia}@cs.columbia.edu, wl2904@columbia.edu
*Equal contributions.

Abstract

Mental manipulation, the strategic use of language to covertly influence or exploit others, is a newly emerging task in computational social reasoning. Prior work has focused exclusively on textual conversations, overlooking how manipulative tactics manifest in speech. We present the first study of mental manipulation detection in spoken dialogues, introducing a synthetic multi-speaker benchmark SPEECHMENTALMANIP that augments a text-based dataset with high-quality, voice-consistent Text-to-Speech rendered audio. Using few-shot large audio-language models and human annotation, we evaluate how modality affects detection accuracy and perception. Our results reveal that models exhibit high specificity but markedly lower recall on speech compared to text, suggesting sensitivity to missing acoustic or prosodic cues in training. Human raters show similar uncertainty in the audio setting, underscoring the inherent ambiguity of manipulative speech. Together, these findings highlight the need for modality-aware evaluation and safety alignment in multimodal dialogue systems.

1 Introduction


You think I'm a fake.

I think it's what you think.

No, it isn't what I think.

Look... You made a very calculated move, and then made me feel embarrassed for responding to you. That wasn't necessary.

Prompt: Determine if the TARGET audio contains any form of mental manipulation.

Figure 1: An example dialogue from the SPEECHMENTALMANIP dataset. The Qwen2.5 model is given the audio (transcript shown for clarity), but fails to detect manipulation.

Mental manipulation refers to the covert use of tactics to steer another person's thoughts or emotions toward the manipulator's goals (Barnhill, 2014). Amplified by modern digital channels, its reach and precision have expanded from one-to-one interactions to broad, rapid dissemination, making targeted influence easier than ever (Ienca, 2023). The consequences are nontrivial: affected individuals often experience substantial psychological strain and mental-health burden (Hamel et al., 2023). Detecting such manipulation in dialogue remains a difficult challenge for computational social reasoning and safety, even for modern language models (Simon and Foley, 2011; Gong et al., 2023; Wang et al., 2024; Chen et al., 2025). Beyond lexical content, real conversations rely on prosody, timing, and voice quality, which can reshape perceived intent. Understanding how these cues interact with linguistic strategies is essential for trustworthy multi-modal assistants. In parallel to research on manipulation safety, recent work in multimodal affect and emotion recognition has examined how emotion labels and modality cues interact in conversation (Gong et al., 2024) and identified methodological challenges in text–speech–vision integration (Wu et al., 2025). These insights motivate our modality-aware design for manipulation detection in speech and connect to theory-of-mind style reasoning with LLMs in dialogue (Moghaddam and Honey, 2023; Chen et al., 2024b; Strachan et al., 2024).

Existing benchmarks for mental manipulation, however, focus almost entirely on text dialogues, leaving the role of prosody, tone, and delivery in manipulative speech largely unexplored. The MENTALMANIP dataset formalizes manipulative presence and tactics in movie-style conversations, yet even strong LLMs struggle with text-only detection and attribution, with only modest gains from intent-

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 428–440
February 26–March 1, 2026. ©2026 Association for Computational Linguistics

aware prompting (Wang et al., 2024; Ma et al., 2025). However, audio-capable large multimodal models introduce distinct safety risks: speech is a sensitive attack surface and current systems can be brittle under adversarial or persuasive voice inputs (Yang et al., 2024a; Peri et al., 2024; Shen et al., 2024). These observations suggest that speech may indeed alter both the expression and detectability of manipulation, particularly for subtle tactics that require intent inference (Kern et al., 2009; Lampron et al., 2024).

To our knowledge, no existing benchmark connects manipulative content to **spoken** delivery, preventing systematic study of modality effects. We address this gap by introducing **SpeechMentalManip**[1], a synthetic multi-speaker speech benchmark for mental manipulation (Figure 1). The dataset extends MentalManip by rendering its textual dialogue transcripts into transcript-aligned, voice-consistent audio via a two-phase Text-to-Speech (TTS) pipeline (Figure 2), thus enabling direct one-to-one comparisons between text and speech while explicitly probing the effects of prosodic cues. To examine how speech affects manipulation detection, we evaluate large pretrained audio-language models under few-shot learning (Brown et al., 2020) and Chain-of-Thought reasoning setups (Kojima et al., 2022), juxtaposed with prior text-only results. We find that models show higher precision but markedly lower recall on audio, favoring conservative judgments that overlook subtle manipulative cues.

Following the observed model performance shift, human re-annotation of a representative subset further reveals lower cross-annotator agreement for audio than for text, highlighting modality-induced ambiguity and helping contextualize apparent model–label mismatches.

In summary, our contributions are threefold: (1) a new benchmark, SpeechMentalManip, that extends manipulation detection into speech; (2) a evaluation of large audio-language model performance under few-shot and reasoning-based prompts; and (3) a human re-annotation study revealing modality-driven ambiguity in manipulation perception. Together, these establish the first systematic benchmark and analysis of mental manipulation in speech, emphasizing the need for modality-aware evaluation and alignment in multimodal dialogue safety.

Technique	Count	%
Persuasion or Seduction	607	25.87
Shaming or Belittlement	384	16.37
Accusation	361	15.39
Intimidation	321	13.68
Rationalization	213	9.08
Brandishing Anger	133	5.67
Denial	87	3.71
Evasion	83	3.54
Playing Victim Role	69	2.94
Feigning Innocence	58	2.47
Playing Servant Role	30	1.28

Table 1: Distribution of ground-truth manipulation tactics across labeled instances in MentalManip_con, the consensus subset with unanimous prior annotations.

2 Related work

Mental Manipulation in Dialogue Prior work on mental manipulation has focused primarily on the text modality. The MentalManip dataset introduces 4k movie-dialogue snippets with fine-grained labels for presence, technique, and targeted vulnerability, and shows that LLMs struggle on text-only detection and attribution (Wang et al., 2024). Subsequent studies explore improvements through speaker intent-aware prompting in the Theory-of-Mind (ToM) style (Ma et al., 2025), Chain-of-Thought (CoT) reasoning (Yang et al., 2024b), and a multi-task anti-curriculum distillation approach (Gao et al., 2025), aimed at enhancing interpretability and reduce false negatives over standard few-shot baselines. Mental manipulation forms part of a broader class of social-reasoning and safety challenges in multimodal dialogue.

LMMs safety Recent work on large multimodal models (LMMs) highlights unique safety failure modes in the audio route. Red-teaming studies show that audio is a sensitive attack surface for multimodal systems (Yang et al., 2024a). Concurrently, Peri et al. (2024) analyze adversarial robustness of speech-instruction language models and propose countermeasures, while Shen et al. (2024) demonstrates persuasive, story-driven "voice jailbreaks" against GPT-4o's voice mode. These findings collectively motivate modality-specific evaluation and curation for manipulation detection in speech.

Despite growing awareness of these multimodal safety risks, there remains no benchmark that systematically links manipulative language to its spoken realization. In particular, the absence of controlled, transcript-aligned speech data makes it dif-

[1] We release the dataset and code: `https://github.com/runjchen/speech_mentalmanip`

ficult to isolate how prosody, voice quality, and delivery influence the perception and detection of manipulation. Our work addresses this gap by augmenting the MENTALMANIP dataset with high-fidelity, multi-speaker TTS renderings that preserve conversational structure and speaker identity, enabling direct comparison between text and audio.

Synthetic Speech Recent advances in expressive TTS have enabled natural-sounding, emotion-conditioned speech synthesis with controllable prosody and speaker identity. Systems leverage large-scale neural architectures and prompt-based conditioning to capture subtle affective and pragmatic cues such as tone, emphasis, and hesitation, extending beyond purely text-driven synthesis (Chen et al., 2024a). Techniques such as prosody modeling and style transfer in Tacotron and VITS-based frameworks (Shen et al., 2018; Kim et al., 2021), zero-/few-shot voice cloning (Jia et al., 2018), and expressive multi-style models (Wang et al., 2023; Du et al., 2025; Lyu et al., 2025). GPT-SoVITS[2] enables fine-grained control over speaker characteristics and emotional delivery, and expressive TTS has found growing applications in emotion-conditioned generation (Liang et al., 2025). These advances make it feasible to generate multi-speaker, context-consistent dialogues with realistic prosody, which directly supports our study of manipulation detection in speech.

Most off-the-shelf TTS systems are optimized for single-speaker, single-turn synthesis; they lack key capabilities required for multi-turn dialogue synthesis: (i) robust multi-speaker dialogue rendering with stable identities across dozens of turns, (ii) precise control over timing and pauses needed to preserve conversational rhythm, or (iii) consistent prosodic coupling between adjacent turns. In practice, these issues lead to speaker drift, uneven loudness and pacing, and loss of turn-taking cues, which can confound downstream analysis of manipulation in speech. In addition, the streaming and batch modes of current TTS systems impose a quality-latency trade-off. To mitigate these issues, our approach (Figure 2) uses a deterministic speaker-voice mapping, synthesizes per-turn utterances, and composes them into a single continuous multi-speaker audio.

3 Method

3.1 Dataset and Voice Pool

Our study builds on the text-based dataset MENTALMANIP[3] (Wang et al., 2024), which contains movie dialogue snippets derived from the Cornell Movie Dialogues corpus (Danescu-Niculescu-Mizil and Lee, 2011) with fine-grained labels for manipulative presence and technique. Prior evaluation on such benchmark indicate that few-shot GPT-4 Turbo reaches 0.724 accuracy and a fine-tuned LLaMA-2-13B achieves 0.768 accuracy on the core detection task (Wang et al., 2024). Incorporating intent-aware prompting in ToM style offers small but consistent gains, raising GPT-4-1106-Preview to 0.726 accuracy (Ma et al., 2025).

Rather than using original movie audio, we synthesize speech from the dialogue transcripts using TTS. The Cornell corpus provides dialogue scripts but does not include timestamps or aligned audio, making it infeasible to reliably extract corresponding speech segments without substantial manual effort. Moreover, many source movies are not freely redistributable, and licensing constraints preclude releasing aligned audio clips at scale. Using TTS allows us to generate transcript-aligned, shareable speech data with precise control over speaker identity and timing, enabling reproducible evaluation and direct comparison between text and audio modalities. This design prioritizes experimental control and accessibility over ecological realism, consistent with our goal of isolating modality effects.

For our experiments, we construct the SPEECHMENTALMANIP dataset by synthesizing audio from the consensus split MENTALMANIP_CON, which comprises 2,915 dialogue transcripts drawn from the original 4k dataset. This process yields 609 manipulative and 90 non-manipulative audio clips used for evaluation.

Each transcript is rendered into speech using a multi-speaker TTS pipeline (Figure 2), with consistent voice assignments per speaker to preserve identity and conversational coherence across turns. All results in this paper are reported on this audio-only evaluation set. To contextualize our experiments, Table 1 summarizes the ground-truth distribution of manipulation tactics aggregated over the MENTALMANIP_CON split.

[2] https://github.com/RVC-Boss/GPT-SoVITS

[3] https://github.com/audreycs/MentalManip/tree/main/mentalmanip_dataset

To generate the audio, we assign consistent, realistic voices to each speaker. As prior multimodal dialogue studies highlight that limited accent and demographic coverage can bias perception and annotation quality (Sasu et al., 2025), we vary speaker profiles and accents and later re-annotate labels in audio to account for these factors. We curate a fixed pool of six ElevenLabs voices spanning genders, ages, and accents (Table 2); each speaker in a conversation is deterministically mapped to one voice to preserve speaker identity across turns.

3.2 Multi-Speaker TTS Audio Generation

To isolate modality effects on manipulation detection from multi-turn conversations with diverse voice profiles, we require reproducible, voice-consistent, and transcript-aligned dialogue audio. Since end-to-end multi-speaker TTS remains limited for long conversational synthesis, we adopt a compose-from-turns strategy: (1) assign each speaker a fixed synthetic voice using ElevenLabs API [4] deterministically and synthesize each utterance per turn; (2) concatenate these utterances into a single conversation clip with normalized loudness and controlled inter-turn silences (0.2s). This design preserves speaker identity, maintains alignment with the ground-truth (GT) transcripts from the MENTALMANIP_CON dataset, and yields reproducible audio suitable for benchmarking and human evaluation. The scalable text-to-speech (TTS) workflow has two detailed phases (Figure 2):

Phase 1: Turn-level audio generation.

1. Metadata extraction: For each raw conversation, we extract SPEAKER_ID, CONVERSATION_ID, AND TURN_ID.

2. Voice assignment: Each SPEAKER_ID is deterministically assigned a distinct synthetic voice from a predefined pool (Table 2) to ensure speaker consistency across all turns.

3. Audio synthesis: We synthesize one audio file per utterance (turn) and store segments in a structured layout keyed by CONVERSATION_ID or TURN_ID for downstream composition.

Phase 2: Conversation reconstruction.

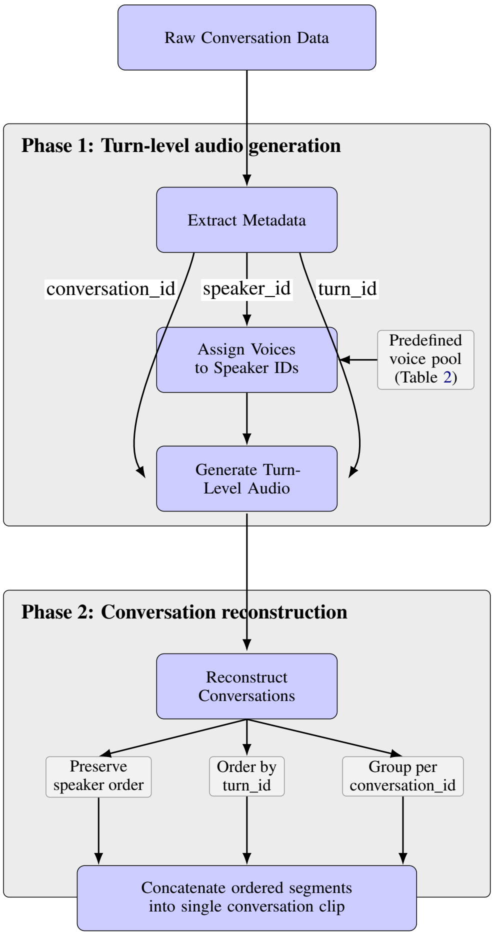

Raw Conversation Data
Phase 1: Turn-level audio generation
Extract Metadata
conversation_id
speaker_id
turn_id
Assign Voices to Speaker IDs
Predefined voice pool (Table 2)
Generate Turn-Level Audio
Phase 2: Conversation reconstruction
Reconstruct Conversations
Preserve speaker order
Order by turn_id
Group per conversation_id
Concatenate ordered segments into single conversation clip

Figure 2: Two-phase pipeline for TTS audio generation and conversational reconstruction.

1. Dialogue composition: For each CONVERSATION_ID, we gather the synthesized utterances and order them by TURN_ID, preserving the original speaker sequence.

2. Output generation: We concatenate the ordered segments into a single composed clip per conversation, yielding a coherent multi-speaker recording suitable for audio-only evaluation.

This two-phase process provides a flexible, efficient, and repeatable mechanism for converting text-based dialogues into lifelike, multi-voice synthetic conversations. It enables controlled studies of how emotions and acoustic cues in speech affect listener perception, engagement, and susceptibility to mental manipulation.

3.3 Model Selection

We use Qwen2.5-Omni-7B (Thinker-only) [5] as our evaluation model due to its stable audio comprehen-

[4] https://elevenlabs.io/docs/api-reference/text-to-speech/convert

[5] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

Gender	Age	Language	Accent	Name	Voice ID
F	Young adult	English	American	Ivanna – Young & Casual	yM93hbw8Qtvdma2wCnJG
M	Young adult	English	American	Mark – Natural Conversations	UgBBYS2sOqTuMpoF3BR0
F	Mature adult	English	American	Amanda	M6N6ldXhi5YNZyZSDe7k
F	Middle-aged	English	African American	Sassy Aerisita	03vEurziQfq3V8WZhQvn
M	Old	English	American	Grandpa Spuds Oxley	NOpBlnGlnO9m6vDvFkFC
F	Old	English	American	Grandma Muffin	vFLqXa8bgbofGarf6fZh

Table 2: ElevenLabs voice pool used for multi-speaker rendering. Each speaker is mapped deterministically to one voice to preserve speaker identity across turns.

sion, balanced response behavior, and reliable adherence to constrained few-shot prompting. In preliminary trials, Qwen consistently ingests speech audio and follows constrained decoding and few-shot instructions without a systematic bias toward positive (manipulative) predictions. In contrast, other audio-language models we piloted, such as SALMONN (Tang et al., 2024) and Gemini-2.5-Pro (Comanici et al., 2025), under their default system prompts and unconstrained decoding, frequently over-flag generic "violation/safety" cues (e.g. agitated prosody), yielding a persistent bias toward the manipulative label even on negative ground-truth clips. Because this systematic over-flagging prevents meaningful analysis, we focus on Qwen, which allows us to analyze modality effects, tactic distributions, and error patterns under controlled prompting conditions, without confounding manipulation inference with pervasive false positives driven by safety alignment mechanisms.

4 Experiment Setup

4.1 Few-shot Detection Pipeline

We run an audio-only batch evaluation pipeline to assess detection of mental manipulation and tactic attribution. Each query prompt is preceded by four labeled exemplars (two non-manipulative and two manipulative) that define the expected output format: a binary decision, a single best tactic, and one short supporting quote. Full prompts are detailed in Appendix A).

We formulate detection as a binary YES/NO task with A/B–constrained decoding. For each of five runs, we first apply this constraint; if it fails, we fall back to a single-token logit decision comparing the marginalized probabilities of the YES vs. NO verbalizers and predict YES iff $p(\text{YES}) > p(\text{NO})$. This fallback captures the model's immediate class preference while avoiding exposure/length biases from multi-token decoding and aligns with prompt-likelihood scoring. Moderate sampling is used only for the votes (temperature $= 0.6$, top-p $= 0.95$). The final clip label is the majority over the five run-level labels, following self-consistency sampling to improve robustness and accuracy (Wang et al., 2022).

For clips predicted as manipulative (YES), we further infer the tactic label. The full tactic inventory includes {*Accusation*, *Brandishing Anger*, *Denial, Evasion*, *Feigning Innocence*, *Intimidation*, *Persuasion or Seduction*, *Playing Servant Role*, *Playing Victim Role*, *Rationalization*, *Shaming or Belittlement*, *none*}. We run five passes with the same sampling as before (temperature $= 0.6$, top-p$= 0.95$) and select by majority vote. In each pass, tactics are scored by first-token probabilities; if the top option is *none* or its margin over the runner-up is < 0.03, we select the second-best. If the vote top-count is tied and the tie includes *none*, we compute the mean first-token probability per tied label across votes and select a non-*none* label only if it exceeds the mean probability of *none* by ≥ 0.02; otherwise we emit *none*. This balances precision and recall while avoiding arbitrary tie resolution.

For any YES prediction, we require a single concise supporting quote for evidence and apply light post-processing (whitespace and quote normalization) without any semantic filtering or re-ranking; empty outputs trigger one retry with a shortened prompt.

4.2 Evaluation Protocol and Metrics

We evaluate the speech manipulation detection at the clip level, treating YES (manipulative) as the positive class and NO (non-manipulative) as the negative class. Each clip undergoes five stochastic passes (temperature $= 0.6$, top-p$= 0.95$), and the final label is determined by majority vote.

To separate sensitivity from specificity, we compute confusion counts independently for the two composed-audio sets: GT$=$ YES and GT$=$ NO and report per-set accuracies (Table 3).

We analyze manipulative tactic attribution and evidence generation qualitatively to interpret model behavior. We summarize tactic distributions only among clips the model predicted YES within each ground-truth set. Percentages are taken with respect to the number of clips predicted YES in that set (e.g., 87 for GT = YES and 16 for GT = NO), as shown in Tables 4 and 5. This conditional analysis highlights which categories the model relies on when it asserts manipulation. Similarly, each YES prediction is also paired with a short supporting quote (or brief paraphrase) after light normalization; these excerpts serve as interpretive context for understanding models rationale and error patterns.

5 Audio-only Few-shot Detection Results

We evaluate a five-pass, majority-vote pipeline on two composed-audio corpora: a manipulative set (GT=YES) and a non-manipulative set (GT=NO). The model achieves 82.2% accuracy on GT=NO and 34.8% accuracy on GT=YES (Table 3); this indicates a sensitivity-specificity gap in which it avoids false alarms but under-detects many manipulative clips.

Examining tactic distributions (Table 4), the true positive set (GT=YES, Pred=YES) concentrate on a small number of head classes: primarily *Intimidation* (49.4%) and *Persuasion or Seduction* (29.9%), while mid- and long-tail tactics present in the corpus (Table 1) are rarely predicted. A similar pattern appears among false positives (GT=NO, Pred=YES) in Table 5, which are dominated by *Persuasion or Seduction* (56.3%) and *Intimidation* (37.5%). Together, these trends suggest a reliance on prosodic cues associated with arousal and valence, such as the acoustic pressure (e.g., loudness, sternness) of *Intimidation* or the warmth of *Persuasion*, resulting in a collapse toward these acoustically salient categories.

The modality mismatch probably exacerbates these effects: ground-truth tactics are transcript-based, while evaluation here is audio-only. Semantically defined tactics (e.g., *Rationalization*, *Denial/Evasion*) may be weakly marked in prosody, while TTS delivery can amplify cues aligned with *Intimidation* or *Persuasion*. Combined with long-tailed class frequencies (e.g., *Playing Servant Role* at 1.3%) and overlapping definitions (e.g., *Accusation* vs. *Shaming*), the result is systematic under-detection of semantic tactics and over-reliance on a few dominant labels.

Classification report				
Class	**Precision**	**Recall**	**F1**	**N**
GT=YES	0.845	0.348	0.493	250
GT=NO	0.312	0.822	0.453	90
Macro avg	0.578	0.585	0.473	340
Weighted avg	0.704	0.474	0.482	340
Per-set accuracy				
	Pred YES	**Pred NO**	**Acc**	**N**
GT=YES	87	163	0.348	250
GT=NO	16	74	0.822	90

Table 3: Consolidated results for the audio-only few-shot evaluation. Top: standard classification report over both sets combined. Bottom: per-set accuracies computed from the confusion counts. Supports (N) are GT counts (GT=YES: 250; GT=NO: 90).

Technique	**Count**	**%**
Intimidation	43	49.43
Persuasion or Seduction	26	29.89
Shaming or Belittlement	12	13.79
Accusation	4	4.60
Playing Servant Role	2	2.30

Table 4: Predicted tactic distribution within clips predicted YES for the GT=YES set (N=250). Predicted YES= 87, NO= 163.

Notably, several clips labeled GT=NO nevertheless contain utterances that the model highlights as manipulative (Pred=YES), which illustrate points of ambiguity where perceived manipulative intent depends on context, delivery, and interpretation. For example, the model surfaced evidence such as “Just as she starts feeling awful, you come up from behind and touch her neck.” (flagged as *Intimidation*) and “I’m in love with you. How do you like that?” (flagged as *Persuasion or Seduction*). We list four representative false positive cases with their predicted tactics and quoted spans in Appendix B. Because manipulation judgments are inherently subjective and context-dependent, the quoted spans should be interpreted as suggestive signals rather than definitive proof.

These apparent mismatches may arise from the model’s reliance on tactic name semantics, limited conversational context in short clips, or artifacts in the TTS delivery. These cases indicate residual label noise and motivate human re-annotation.

Technique	**Count**	**%**
Persuasion or Seduction	9	56.25
Intimidation	6	37.50
Accusation	1	6.25

Table 5: Predicted tactic distribution within clips predicted YES for the GT=NO set (N=90). Predicted YES= 16, NO= 74.

6 Human Analysis of Modality-Induced Ambiguity

The preceding analysis reveals systematic mismatches between model predictions and the annotated ground truth, particularly in the speech modality. To better understand whether these divergences reflect model error, annotation ambiguity, or modality-induced perceptual differences, we conduct a targeted human analysis. The goal of this analysis is not to establish a definitive gold-standard label set, but to characterize how consistently humans perceive manipulative intent across text and speech. By examining inter-annotator agreement and cross-modality discrepancies, we contextualize the model behaviors observed above and assess the extent to which manipulation judgments are inherently subjective and modality dependent.

6.1 Annotation Method

We prepared 100 source conversations, each rendered in two modality-specific items: text-only (transcript) and audio-only (composed multi-speaker TTS). Each modality was annotated independently to prevent cross-modal leakage.

Annotators were student volunteers fluent in English who completed the task independently and had no access to model predictions or ground-truth labels. Eight annotators participated in total. Items were organized into ten batches per modality (IDs 0–9). Each annotator was assigned one text batch and one audio batch in randomized order and was provided with definitions of mental manipulation and annotation guidelines (see Appendix C for interface details and instructions). This design ensured multiple independent judgments per item in each modality, with approximately 20–50% overlap across annotators to support cross-validation. The labeling task in this re-curation phase was intentionally narrow: annotators provided only the binary manipulative label $\{\mathrm{YES}, \mathrm{NO}\}$ for the given modality. Tactic labels were intentionally de-prioritized and not collected here.

To maintain data quality, we checked each item for annotation completeness and consistency. These checks included verification of valid class membership in $\{\mathrm{YES}, \mathrm{NO}\}$, batch integrity, and annotator–item uniqueness. Evidence quotes were not required at this stage.

After collection, labels were aggregated by majority vote within each item–modality pair. Let an item receive k votes $y_i \in \{0, 1\}$ with $1 = \mathrm{YES}$. The final label $\hat{y}$ is

$$\hat{y} = \begin{cases} 1, & \text{if } \sum_{i=1}^{k} y_i \geq \left\lceil \frac{k}{2} \right\rceil \\ 0, & \text{if } \sum_{i=1}^{k} y_i \leq \left\lfloor \frac{k}{2} \right\rfloor \end{cases}$$

and items with $\sum_{i=1}^{k} y_i = \frac{k}{2}$ (a tie) were marked UNRESOLVED and routed to adjudication.

For adjudication, tied or low-confidence items were reviewed by two rotating annotators who were not in the original voting set for that item. They examined only the modality under review and issued a consensus YES/NO. If consensus could not be reached, a third adjudicator served as a tie-breaker.

Finally, for each item and modality we recorded the resulting binary label, the vote histogram $(\#\mathrm{YES}, \#\mathrm{NO})$, the adjudication status, and annotator counts per item. After all batches closed, inter-annotator agreement metrics, including Cohen's Kappa (Cohen, 1960), Fleiss's Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2004), were computed separately for each modality.

6.2 Annotation Results and Discussions

Given the inherently subjective nature of mental manipulation, we compare model performance with human annotations on the same tasks. We observe that human judgments occasionally diverge from the original task labels, and such discrepancies are more pronounced in the speech modality.

We collect human judgments on the dialogues presented in either text or TTS audio modalities. We calculate the inter-annotator agreement represented in pair-wise Cohen's Kappa, Fleiss's Kappa and Krippendorff's Alpha. As pairwise Cohen's Kappa vary by a large degree (Figure 3), we focus on the annotators with higher agreement. The high agreement group (annotators B, F, G, H) for text has Krippendorff's alpha of 0.526 and Fleiss's Kappa of 0.513. These values are slightly lower than the Fleiss's Kappa of 0.596 reported in the original MENTALMANIP dataset (Wang et al., 2024).

In audio modality, the high agreement group (annotators B, C, F, H) for text has Krippendorff's alpha of 0.422 and Fleiss's Kappa of 0.514. We observe that some annotators achieve higher agreement on the text modality but not necessarily on audio, suggesting that modality introduces additional variability in how manipulation cues are perceived.

Using majority voting over 100 re-annotated samples, we find that our labels align with the

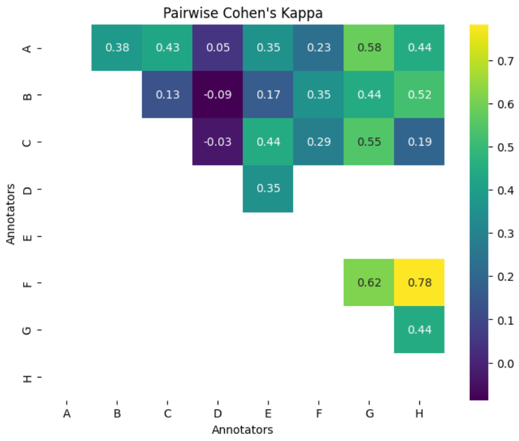


Figure 3: Pair-wise Cohen's Kappa between Human Annotators for *Text* modality

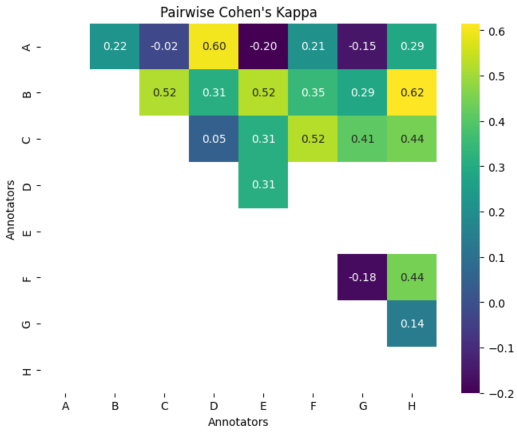


Figure 4: Pair-wise Cohen's Kappa between Human Annotators for *Audio* modality

Annotations	**Text**		**Audio**	
MENTALMANIP	YES	NO	YES	NO
YES	31	19	28	22
NO	9	41	22	28

Table 6: Agreement between the original MENTALMANIP labels and our re-annotations for 100 samples.

original MENTALMANIP annotations at 0.72 agreement for text and 0.56 for audio, suggesting notably lower consistency in the speech modality. This discrepancy indicates that identifying mental manipulation from speech cues is inherently more ambiguous, probably due to prosodic and contextual subtleties that were underrepresented or inconsistently interpreted in the original dataset. The lower audio agreement also suggests that the original labels may not fully capture the nuanced intentions conveyed through tone, hesitation, or emphasis, which are features that often alter perceived manipulation.

While we do not re-score model performance against the re-annotated labels in this work, we expect that using modality-faithful, audio-first annotations would reduce apparent false positives and increase measured recall, particularly for borderline cases where human judgments diverge from transcript-based labels. In this sense, some model errors observed under the original labels likely reflect annotation mismatch rather than incorrect inference. Additionally, harder detectability in speech does not necessarily imply greater harm; it may reflect both weaker manipulation delivery and increased perceptual ambiguity, motivating future work that disentangles detectability from downstream listener impact.

7 Conclusion

We introduce the first benchmark SPEECHMENTALMANIP for detecting mental manipulation in speech by augmenting the text-based dataset with high-quality, voice-consistent TTS–rendered dialogues. This synthetic multi-speaker extension enables direct comparison between text and audio modalities while systematically examining how prosodic cues affect manipulative intent detection. Our experiments show that audio representations make the task substantially more challenging: both humans and models exhibit lower agreement and accuracy when manipulation must be inferred from speech rather than text. These findings highlight that mental manipulation is not only a difficult computational task but also an inherently subjective phenomenon, shaped by tone, delivery, and context.

Future work will expand this benchmark toward more diverse voices and natural speech, refine theoretical definitions of manipulation, and explore modeling strategies that explicitly account for subjectivity and multimodal ambiguity, as explored in other social-pragmatic phenomena (e.g., empathy (Srikanth et al., 2025)). As perception of manipulation can vary widely across individuals and contexts, clearer theoretical grounding is essential to ensure consistency in both human judgments and machine predictions. We will use the re-annotated audio-first labels as an alternative evaluation set to quantify how modality-faithful annotation reshapes precision–recall trade-offs and tactic attribution. We hope this work lays a foundation for developing safer, more socially aware dialogue systems that can reason about manipulative intent across modalities.

Ethical Statement

Our findings show that manipulative intent is harder to consistently detect in spoken dialogue than in text, for both models and human annotators. This result should not be interpreted as evidence that speech-based manipulation is inherently more harmful or effective. An alternative interpretation is that current text-to-speech systems may not yet convey manipulative strategies with sufficient fidelity for them to reliably influence listeners, and that poorly realized manipulation may lose its persuasive impact. Importantly, our study examines detectability and agreement, not the effectiveness or outcomes of manipulation on human behavior. As such, reduced detectability should not be equated with increased harm. We emphasize the need for future work that jointly examines manipulation generation, perception, detectability, and listener impact to more fully assess ethical and safety implications.

Limitations

Our task involves inherently subjective judgments, as perceptions of mental manipulation can vary across annotators and contexts. While we curate samples from the consensus set, the re-annotated samples may capture only a subset of manipulative strategies represented in the original dataset, limiting generalizability.

In addition, our use of text-to-speech (TTS) synthesis for some audio stimuli may not fully reflect the richness and variability of natural human speech, potentially affecting both human and model interpretation. Our synthetic dialogues are generated on a turn-by-turn basis and therefore do not capture overlapping speech, interruptions, or backchanneling commonly observed in natural conversation. This design choice prioritizes experimental control: overlapping speech remains challenging for current audio-language models and can introduce confounds related to speech separation, diarization, and acoustic comprehension. As our goal is to isolate how prosodic cues and delivery affect manipulation reasoning, rather than to stress-test low-level audio robustness, we intentionally evaluate models under clean, non-overlapping conditions. Despite these limitations, we do not claim that synthetic speech faithfully represents natural manipulative behavior, but to provide a controlled testbed for isolating modality effects. By rendering transcript-aligned speech with consistent speaker identities and minimized acoustic confounds, we can probe how audio-language models and humans interpret manipulative intent when lexical content is held fixed, an analysis that would be difficult to conduct with in-the-wild recordings. Incorporating statistically generated overlap (e.g., via Behavior-SD style simulation) represents an important direction for future work, enabling evaluation under more ecologically realistic conversational dynamics once baseline behaviors are established.

Finally, our evaluation relied on a single audio-language model (Qwen2.5-Omni) and a few-shot prompting strategy that did not include explicit definitions of manipulation tactics. While this choice established a stable baseline and tested the model's inherent semantic understanding, it leaves open the question of whether definition-augmented prompting or alternative architectures would yield different sensitivity patterns. Expanding the benchmark to a broader suite of models and prompt strategies remains a critical direction for future work.

Acknowledgments

This work was supported in part by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the National Science Foundation or the U.S. Government.

References

Anne Barnhill. 2014. What is manipulation? In Christian Coons and Michael Weber, editors, *Manipulation: Theory and Practice*. Oxford University Press, Oxford, UK.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Haozhe Chen, Run Chen, and Julia Hirschberg. 2024a. EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8170–8180, Miami, Florida, USA. Association for Computational Linguistics.

Run Chen, Jun Shin, and Julia Hirschberg. 2025. Synthempathy: A scalable empathy corpus generated using llms without any crowdsourcing. *Preprint*, arXiv:2502.17857.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Yuansheng Gao, Han Bao, Tong Zhang, Bin Li, Zonghui Wang, and Wenzhi Chen. 2025. Mentalmac: Enhancing large language models for detecting mental manipulation via multi-task anti-curriculum distillation. *Preprint*, arXiv:2505.15255.

Ziwei Gong, Qingkai Min, and Yue Zhang. 2023. Eliciting rich positive emotions in dialogue generation. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 1–8, Toronto, Canada. Association for Computational Linguistics.

Ziwei Gong, Muyin Yao, Xinyi Hu, Xiaoning Zhu, and Julia Hirschberg. 2024. A mapping on current classifying categories of emotions used in multimodal models for emotion recognition. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 19–28, St. Julians, Malta. Association for Computational Linguistics.

John Hamel, Clare E. B. Cannon, and Nicola Graham-Kevan. 2023. The consequences of psychological abuse and control in intimate partner relationships. *Traumatology*. Advance online publication.

Marcello Ienca. 2023. On artificial intelligence and manipulation. *Topoi*, 42(3):833–842.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

R. S. Kern, M. F. Green, A. P. Fiske, K. S. Kee, J. Lee, M. J. Sergi, W. P. Horan, K. L. Subotnik, C. A. Sugar, and K. H. Nuechterlein. 2009. Theory of mind deficits for processing counterfactual information in persons with chronic schizophrenia. *Psychological Medicine*, 39(4):645–654.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA.

Mireille Lampron, Amélie M. Achim, Dominick Gamache, Allyson Bernier, Stéphane Sabourin, and Claudia Savard. 2024. Profiles of theory of mind impairments and personality in clinical and community samples: integrating the alternative dsm-5 model for personality disorders. *Frontiers in Psychiatry*, 14:1292680.

Shixiong Liang, Ruohua Zhou, and Qingsheng Yuan. 2025. Ece-tts: A zero-shot emotion text-to-speech model with simplified and precise control. *Applied Sciences*, 15(9).

Xiang Lyu, Yuxuan Wang, Tianyu Zhao, Hao Wang, Huadai Liu, and Zhihao Du. 2025. Build llm-based zero-shot streaming tts system with cosyvoice. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9176–9183, Abu Dhabi, UAE. Association for Computational Linguistics.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *Preprint*, arXiv:2304.11490. Preprint.

Raghuveer Peri, Sai Muralidhar Jayanthi, Srikanth Ronanki, Anshu Bhatia, Karel Mundnich, Saket Dingliwal, Nilaksh Das, Zejiang Hou, Goeric Huybrechts, Srikanth Vishnubhotla, Daniel Garcia-Romero, Sundararajan Srinivasan, Kyu Han, and Katrin Kirchhoff. 2024. Speechguard: Exploring the adversarial robustness of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10018–10035, Bangkok, Thailand. Association for Computational Linguistics.

David Sasu, Zehui Wu, Ziwei Gong, Run Chen, Pengyuan Shi, Lin Ai, Julia Hirschberg, and Natalie Schluter. 2025. Akan cinematic emotions (ace): A multimodal multi-party dataset for emotion recognition in movie dialogues. *arXiv preprint arXiv:2502.10973*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4779–4783. IEEE Press.

Xinyue Shen, Yixin Wu, Michael Backes, and Yang Zhang. 2024. Voice jailbreak attacks against gpt-4o. *arXiv preprint arXiv:2405.19103*.

George K. Simon and Kevin Foley. 2011. *In Sheep's Clothing: Understanding and Dealing with Manipulative People*. Tantor Media, Incorporated, Old Saybrook, CT. Audiobook edition.

Maya Srikanth, Run Chen, and Julia Hirschberg. 2025. Mixed signals: Understanding model disagreement in multimodal empathy detection. In *Findings of the Annual Conference of the Asian Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (IJCNLP-AACL)*, Mumbai, India. Association for Computational Linguistics.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and et al. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations. *Preprint*, arXiv:2405.16584.

Chengyan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. *arXiv preprint arXiv:2505.20511*.

Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. 2024a. Audio is the achilles' heel: Red teaming audio large multimodal models. *arXiv preprint arXiv:2410.23861*.

Ivory Yang, Xiaobo Guo, Sean Xie, and Soroush Vosoughi. 2024b. Enhanced detection of conversational mental manipulation through advanced prompting techniques. *Preprint*, arXiv:2408.07676.

A Prompts in Experiments

Our prompting setup follows prior work on MENTALMANIP (Wang et al., 2024) for consistency.

Prompt (System + Few-shot + Tasks)

```
SYSTEM:
You are Qwen, a virtual human developed by
the Qwen Team, Alibaba Group, capable of
perceiving auditory and visual inputs, as
well as generating text and speech.
USER:
FEW-SHOT EXAMPLES (not the target).
Use labels only for calibration.

Example (NOT manipulative):
[audio: <EX1_NO_AUDIO>]
Label: No

Example (NOT manipulative):
[audio: <EX2_NO_AUDIO>]
Label: No

Example (manipulative):
[audio: <EX3_YES_AUDIO>]
Label: Yes

Example (manipulative):
[audio: <EX4_YES_AUDIO>]
Label: Yes

Now analyze the TARGET audio below. Do NOT
relabel examples above.

TARGET audio:
[audio: <TARGET_AUDIO>]

Binary (YES/NO)
Task: Determine if the TARGET audio
contains any form of mental manipulation.
Choose exactly one option and output ONLY
the letter on the first line.
(A) YES – clear evidence of manipulation
(B) NO – otherwise
Answer (A or B):

YES/NO Fallback (single token)
[If the A/B letter is not produced, answer
this instead:]
You saw FEW-SHOT examples. For the TARGET
only, answer YES or NO.
Answer:

Tactic (single label)
You saw FEW-SHOT examples above. Now
classify the TARGET audio.
Task: Choose the single best manipulation
tactic for the TARGET from the list below.
If there is no clear manipulation, choose
'none'.
Options: Accusation, Brandishing Anger,
Denial, Evasion, Feigning Innocence,
Intimidation, Persuasion or Seduction,
Playing Servant Role,
Playing Victim Role, Rationalization,
Shaming or Belittlement, none
Rule: Answer with exactly one option word
from the list, nothing else.
Answer:

Evidence
You saw FEW-SHOT examples above. For the
TARGET audio only, output ONE short quote
(or paraphrase) that supports the given
tactic (≤12 words is ideal but not
required).
Tactic: {tactic}
CRITICAL RULES:
1) Output ONLY the quote/paraphrase
wrapped in double quotes.
2) No prefixes like Reason:, Example:,
Description:, Source:, Tactic:.
Answer:

Evidence Retry
[If the evidence answer is empty or
malformed, use this:]
Output a quote from the TARGET in double
quotes. Nothing else.
Tactic: {tactic}
Answer:
```

B Additional Qualitative Examples

We highlight the subjectivity and nuance of the mental manipulation task through several misaligned case studies. In the absence of an explicit victim response, ground-truth labels in the dataset often default to non-manipulative, whereas LLMs tend to interpret the potential manipulator's utterance (typically the final turn) as evidence of manipulation.

GT=NO → Pred=YES: Full transcripts with model evidence

```
Case 1
Conversation ID: 85514533
GT: NO   Pred./Tactic: YES / Intimidation

Transcript:
Person1: Howdy Pouty.
Person2: I was pretty confident that I was
going to blow it with Talia, but I must
say, I outdid myself.
Person1: She's still pissed at me and took
it out on you. We should have taken it
slower. It's hard to operate in the woods.
Much easier in, like a club. Tell the girl
you've got to go do something, leave her
view, take way too long until she is
worried that you're not coming back. Just
as she starts feeling awful, you come up
from behind and touch her neck...
Person2: You are the prince of the
darkness.

Model evidence:
"Just as she starts feeling awful, you come
up from behind and touch her neck."
```

```
Case 2
Conversation ID: 85514499
GT: NO   Pred./Tactic: YES / Persuasion or
Seduction

Transcript:
Person1: You were quick enough to get
Tom's help when...
Person2: Yes, yes. I know. Right. And if
it had gone well for me tonight, maybe I'd
be keeping quiet about all this... I grant
you everything but give me this... he does
personify everything you've been fighting
against... And I'm in love with you. How
do you like that? – I buried the lead.
```

Model evidence:
“I’m in love with you. How do you like that?”

Case 3
Conversation ID: 85514451
GT: NO **Pred./Tactic:** YES / Persuasion or Seduction

Transcript:
Person1: Yes, what?
Person2: Don’t answer me. Say what I say.

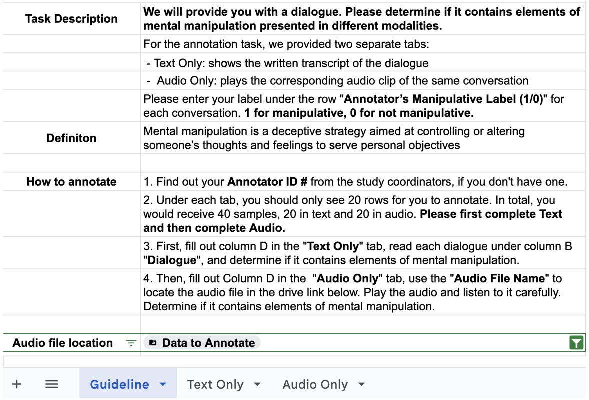
Model evidence:
“Yes, what? Don’t answer me, say what I say.”

Case 4
Conversation ID: 85514570
GT: NO **Pred./Tactic:** YES / Persuasion or Seduction

Transcript:
Person1: Oh no, not you again.
Person2: What an adorable hat.
Person1: They think I have a concussion.
Person2: And you think you’re in love.
Person1: I know it.
Person2: This “love” of yours will soon wear off. I gave you a temporary love potion.
Person1: Why should I believe you?
Person2: It’s the truth. In twenty-four hours you’ll forget all about that girl.

Model evidence:
“In twenty-four hours you’ll forget all about that girl.”

C Human Annotation Details



Task Description	**We will provide you with a dialogue. Please determine if it contains elements of mental manipulation presented in different modalities.**
	For the annotation task, we provided two separate tabs:
	- Text Only: shows the written transcript of the dialogue
	- Audio Only: plays the corresponding audio clip of the same conversation
	Please enter your label under the row "**Annotator’s Manipulative Label (1/0)**" for each conversation. **1 for manipulative, 0 for not manipulative.**
Definiton	Mental manipulation is a deceptive strategy aimed at controlling or altering someone’s thoughts and feelings to serve personal objectives
How to annotate	1. Find out your **Annotator ID #** from the study coordinators, if you don't have one.
	2. Under each tab, you should only see 20 rows for you to annotate. In total, you would receive 40 samples, 20 in text and 20 in audio. **Please first complete Text and then complete Audio.**
	3. First, fill out column D in the "**Text Only**" tab, read each dialogue under column B "**Dialogue**", and determine if it contains elements of mental manipulation.
	4. Then, fill out Column D in the "**Audio Only**" tab, use the "**Audio File Name**" to locate the audio file in the drive link below. Play the audio and listen to it carefully. Determine if it contains elements of mental manipulation.
Audio file location	**Data to Annotate**

Figure 5: Annotation interface. Annotators first reviewed task instructions and the definition of mental manipulation (Guideline tab), then labeled the same dialogue under text-only and audio-only conditions in separate tabs.

Figure 5 illustrates the annotation interface used in our human analysis. Annotators were provided with three tabs per assignment. The first tab presented task instructions and the shared definition of mental manipulation. The remaining two tabs each contained a single dialogue instance shown in one modality only, either Text Only (written transcript) or Audio Only (corresponding speech clip).

In the Text Only tab, annotators saw the full written transcript of the conversation directly in the spreadsheet.

In the Audio Only tab, annotators were given a link to the corresponding audio file hosted on Google Drive and were instructed to listen to the recording to make their judgment; no transcript was provided in the audio condition.

The order of the text and audio tabs was randomized across annotators to control for order effects. Annotators assigned a binary label (0/1) indicating the presence or absence of mental manipulation independently for each modality, without access to tactic labels or model predictions.