CAN YOU HEAR MY MIND? CATCHING THEORY OF MIND IN VOICE AND PRAGMATICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same words can convey different intentions depending on how they are spoken. Interpreting these cues alongside verbal content draws on Theory of Mind (ToM), the capacity to perceive and infer others' mental states. Yet existing ToM benchmarks primarily assess mental-state reasoning from textual or visual events, while audio reserach largely focuses on state recognition rather than cognition. This leaves unclear how vocal delivery and pragmatics shape mind. We introduce Overtone, the first multimodal ToM benchmark to systematically assess how models use vocal and pragmatic cues to infer mental states. Overtone comprises 2,335 social-interaction scenarios with human-recorded speech, enables comprehensive, fine-grained evaluation across diverse classical ToM tasks and a broad spectrum of ToM abilities. A human study and evaluations of 16 speech and omni-modal models reveal that models struggle to maintain coherent mental-state representations and integrate vocal cues into ToM reasoning. Hence, we introduce PEAR (Prediction-Error-Adaptive Reasoning), a training-free ToM reasoning framework inspired by predictive coding in the human brain. Mental-state hypotheses guide the interpretation of multimodal cues, while metacognitive regulation the reasoning process. Extensive experiments show that PEAR improves performance across backbones, outperforms several reasoning baselines, and generalizes across diverse ToM datasets. Our findings highlight current multimodal models' limitations in inferring mental states from vocal and pragmatic cues. Together, our work provides a benchmark and reasoning framework for assessing and improving this capability, advancing multimodal social reasoning.

1 INTRODUCTION

In social interactions, a remarkable capacity of human cognition is to infer others' mental states from limited and ambiguous social information, thus interpreting their behavior and communicative goals (Premack & Woodruff, 1978; Baron-Cohen et al., 1985). This capacity, known as Theory of Mind (ToM), draws on behavioral and expressive information across modalities. As shown in Figure 1, mental-state inference in spoken interaction not only depends on what is said, but also on how it is said. Vocal expression can alter the social meaning of identical words (Chevallier et al., 2011), while context and pragmatics shape its interpretation (Tatár et al., 2026). As multimodal models increasingly process speech directly, a fundamental question arises: **Can they go beyond recognizing spoken content to infer the speaker's mental states through voice and pragmatics?**

Despite broad task coverage (Chen et al., 2024; Jin et al., 2024; Shi et al., 2025), machine ToM benchmarks remain centered on text and visual events, with limited attention to vocal and pragmatic cues. Prosody and other paralinguistic features convey social meaning beyond transcripts (Ma et al., 2026). However, existing evaluations largely remain focused on recognizing basic emotions, vocal attributes, or isolated pragmatic phenomena (Li et al., 2025), focusing primarily on the identification and classification of overt behavioral manifestations (Bird & Viding, 2014). Recognizing such person-level cues does not establish whether models understand social emotions and higher-order pragmatic attitudes, which may require representing beliefs, appraisals, and attitudes toward emotions or others' thoughts (Burnett et al., 2009; Norman et al., 2016). Research in cognitive science further links incongruence between emotional expression and context to greater engagement of mentalizing networks (Biró et al., 2025). More recent studies show that audio LLMs remain strongly biased toward lexical semantics when vocal expression and verbal content con-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

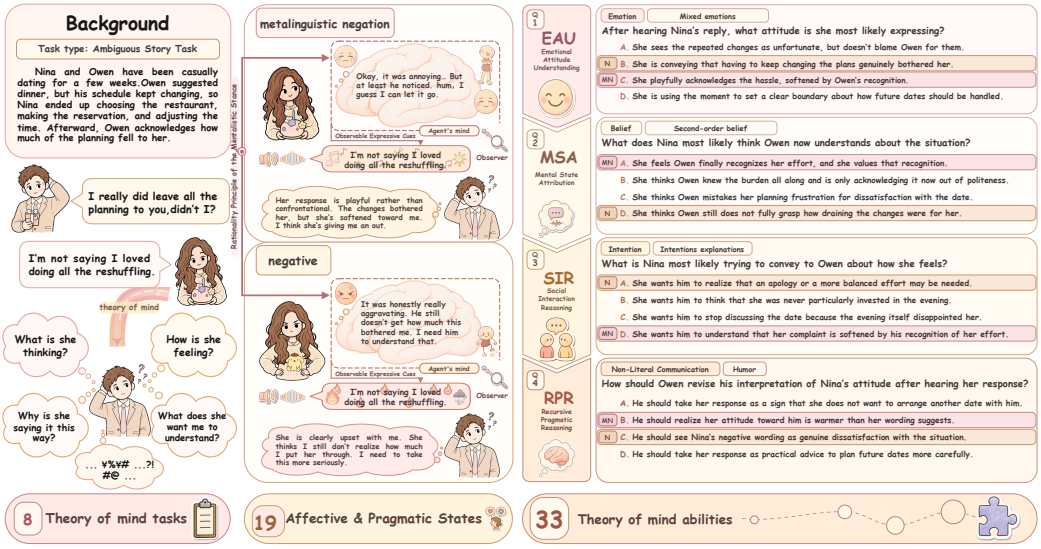

Figure 1: The Overtone framework & examples. The same word in a shared social background can support distinct mental-state interpretations depending on vocal delivery and pragmatics. The example contrasts metalinguistic negation (MN) and negative expression (N).

flict (Chen et al., 2026; Corrêa et al., 2026). **A more diagnostic question is: Can changes in vocal expression lead the model to form different yet correct mental-state interpretations?**

To address these gaps, we introduce **Overtone**, the first human-recorded, multimodal, multi-level ToM benchmark targeting vocal expression, pragmatics, and mental-state reasoning, as shown in Figure 1. All tasks use semantically neutral background and utterances, while shared scripts hold text constant and vary vocal delivery, enabling controlled analysis of how vocal and pragmatic information affects ToM inference. Human studies and evaluations of 16 advanced speech and omni-modal models on Overtone reveal weaknesses in mental-state consistency, vocal-cue utilization, and multimodal integration. Different prompting strategies yields inconsistent gains and can even degrade correct predictions on Overtone, highlighting reliable multimodal integration as a key challenge for ToM reasoning. These findings motivate a central question: **How can models integrate multimodal social information for ToM reasoning in a manner closer to the human brain?** We further introduce **PEAR** (**P**rediction-**E**rror-**A**daptive **R**easoning), a training-free, brain-inspired predictive-coding framework that combines top-down prediction, bottom-up error feedback, and metacognitive regulation to model mental-state formation and revision. Extensive experiments show that PEAR delivers consistent gains across multiple backbones and outperforms several explicit reasoning baselines. More importantly, these gains persist across benchmarks and ToM task settings, demonstrating strong generalization and broad transferability beyond the original benchmark.

Overall, our study addresses gaps across voice, pragmatics, ToM evaluation and multimodal ToM reasoning; advancing social understanding from surface-level state recognition toward cognitive attribution (Baker et al., 2017; Houlihan et al., 2023). Our contributions are threefold: **(i)** The first human-recorded ToM benchmark dedicated to voice and pragmatics, covering 8 ToM tasks, 4 Affective–Pragmatic dimensions with 19 fine-grained types, and 7 ToM ability dimensions with 33 subabilities, structured into four progressive social-cognitive reasoning levels. **(ii)** We establish a human baseline and systematically evaluate advanced models and reasoning strategies, providing a unified empirical characterization of current multimodal ToM capabilities and limitations. **(iii)** We propose PEAR, a training-free ToM reasoning framework that combines brain-inspired predictive coding with metacognitive regulation, demonstrating consistent improvements across model backbones and multimodal ToM tasks, advancing multimodal AI toward more reliable, context-sensitive, and cognitively grounded social understanding of human mental states and interactions.

Table 1: **Comparison with related benchmarks.** ToM Ability lists covered mental-state dimensions using initials. NVV: non-verbal vocalization; Pairs: controlled variations; VC: vocal cues; VCTR/VPTR: vocal-cued/vocal-pragmatic ToM reasoning.

Benchmark	#Sample	ToM Ability	Multi-level QA	NVV	Pairs	VCTR	VC	VPTR
TOMI(Le et al., 2019)	5,994	B	✗	✗	✗	✗	✗	✗
HI-ToM(Wu et al., 2023)	1,200	B	✓	✗	✗	✗	✗	✗
ToMBench(Chen et al., 2024)	2,860	E / D / I / K / B / NLC	✗	✗	✗	✗	✗	✗
MMToM-QA(Jin et al., 2024)	600	B / G	✗	✗	✗	✗	✗	✗
MuMA-ToM(Shi et al., 2025)	900	I / B	✗	✗	✗	✗	✗	✗
MoMentS(Villa-Cueva et al., 2025)	2,344	E / D / I / K / B / NLC / P	✗	✗	✗	✗	✗	✗
StressTest(Yosha et al., 2026)	218	—	✗	✗	✓	✗	✓	✗
CP-Bench(Wang et al., 2025)	2,747	—	✗	✗	✗	✗	✓	✗
MMSU(Wang et al., 2026)	5,000	—	✗	✓	✗	✗	✓	✗
EchoMind(Zhou et al., 2026)	9,323	—	✓	✓	✓	✗	✓	✗
Overtone (Ours)	**10,656**	**E / D / I / K / B / NLC / P**	✓	✓	✓	✓	✓	✓

2 Related Work

Machine Theory of Mind. ToM benchmarks have expanded from textual narratives to multimodal scenarios. ToMi (Le et al., 2019) and Hi-ToM (Wu et al., 2023) assess false-belief reasoning and higher-order belief tracking in text, while ToMBench (Chen et al., 2024) covers eight ToM task types. MMToM-QA (Jin et al., 2024) and MuMA-ToM (Shi et al., 2025) evaluate beliefs and goals through video and text, with the latter covering multi-agent interactions. MoMentS (Villa-Cueva et al., 2025) assesses seven mental-state dimensions using naturalistic short films. For ToM reasoning, SimToM (Wilf et al., 2024) uses perspective-taking, while chain-of-thought prompting supports stepwise inference (Rahimi Moghaddam & Honey, 2023). ToMnet (Rabinowitz et al., 2018) meta-learns agent-specific representations, and SymbolicToM (Sclar et al., 2023) tracks nested beliefs with symbolic graphs. BIP-ALM and LIMP combine language models with inverse planning (Jin et al., 2024; Shi et al., 2025); ThoughtTracing (Kim et al., 2025) propagates weighted mental-state hypotheses, and AutoToM (Zhang et al., 2025) automates agent-model construction and refinement. However, MoMentS reports inconsistent gains from audio over transcripts (Villa-Cueva et al., 2025), while AutoToM converts multimodal observations into text before inference (Zhang et al., 2025).

Vocal Expression and Pragmatic Inference. Vocal expression conveys emotion and social attitudes while shaping how utterances are understood beyond their literal content (Ward & Marco, 2024). Qian & Skantze (2026) examine how lexical form and prosody contribute to backchannel interpretation in conversational context, while StressTest (Yosha et al., 2026) demonstrates that sentence stress alone can alter implied meaning and intent. In their survey of pragmatic evaluation, Ma et al. (2025) emphasize context's role in interpreting implicit meaning, emotion, and communicative intent. Recent benchmarks examine these aspects: CP-Bench (Wang et al., 2025) and MMSU (Wang et al., 2026) evaluate the integration of linguistic content with emotion and prosody, while EchoMind (Zhou et al., 2026) explores vocal expression in socially grounded and empathetic understanding. Table 1 compares Overtone with existing benchmarks, which emphasize acoustic attributes, emotions, and specific pragmatic phenomena over mental-state inference from vocal cues.

3 Overtone Benchmark

3.1 Overview

Overtone comprises 2,335 scripts, 2,664 human-recorded audio clips, and 10,656 questions. Each script pairs a social-interaction context with a target utterance; some also include multi-party dialogue. Both the background and the utterance are neutrally to avoid explicit cues to the target emotion or mental state. For a subset of scripts, we hold the text fixed while varying vocal expression to disentangle its contribution to mental-state reasoning. As shown in Figure 2(a), the scenario is organized along two axes: four Affective–Pragmatics Dimensions covering 19 fine-grained types, and eight psychology-based ToM Task Dimensions. Each audio is paired with four progressively structured questions, each targeting one of seven Ability Dimensions with 33 subabilities (Figure 2(b)).

Following (Chevallier et al., 2011), we distinguish vocal expressions by the degree of mental-state inference involved in their interpretation. Basic emotions can often be recognized from vocal patterns with less reasoning about a speaker's mental states, whereas social emotions and higher-order pragmatic attitudes require perspective-taking, social evaluation, and mental-state attribution. Ac-

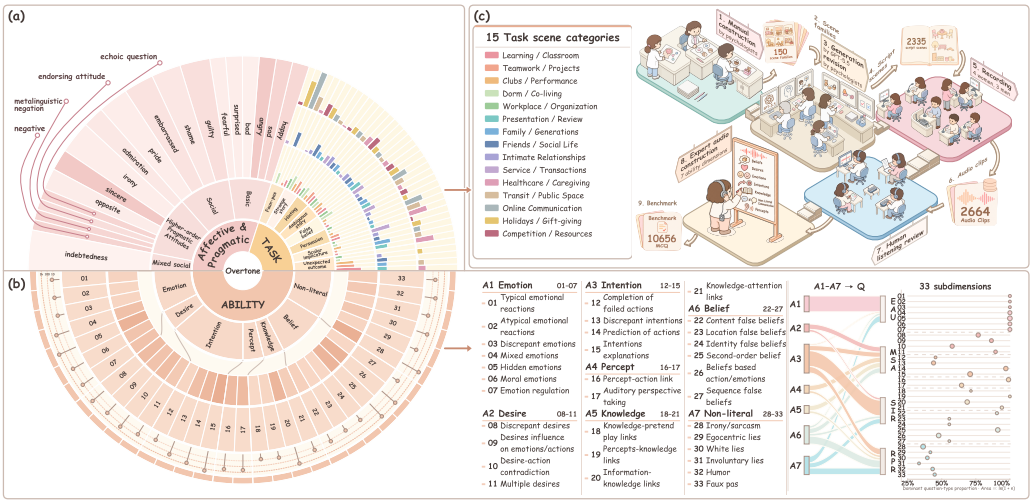

Figure 2: Design framework and construction pipeline of Overtone.

cordingly, Overtone organizes vocal expressions into four categories: Basic Emotion, Social Emotion, Higher-order Pragmatic Attitudes, and Mixed Social Emotion, covering 19 fine-grained types. Importantly, within Basic Emotion, we pair semantically neutral utterances with ambiguous social contexts to evaluate emotion-related mental-state inference rather than emotion recognition alone. Overtone further organizes ToM evaluation along task and ability dimensions, grounded in established psychological ToM taxonomies (Beaudoin et al., 2020). Detailed definitions, theoretical motivations, distributions, and examples are provided in Appendices B.1, B.2, and B.3.

3.2 Dataset Construction & Evaluation Metrics

As shown in Figure 2(c), Overtone is constructed in four stages: social-scenario design, script development, human recording, and multi-level ToM question construction. Psychological materials provide limited spoken stimuli, while multimodal datasets entangle mental-state cues with linguistic, visual, and narrative information. We develop stimuli grounded in psychology and social-cognition research, with eight experts in psychology and linguistics contributing to design and review.

Scenario Library. We define 15 broad social settings and refine them into 150 fine-grained scenarios spanning different relationships, interaction goals, event types, and social conflicts. **Script Construction.** For each scenario, experts specify variations in events, relationships, information access, latent mental states, interaction goals, and target utterances. GPT-5.5 then generates candidate scripts under these constraints, followed by expert review and scoring; detailes are provided in Appendix C.1. Each script contains a social background and a target utterance, and is assigned to one of eight ToM tasks and one of 19 affective–pragmatic types. We also construct **shared scripts**, keeping the text fixed while varying vocal expression. After psychology experts review, the final set contains 2,335 scripts, including 2,049 single-realization scripts and 286 shared-script groups, yielding 2,664 audio clips. **Questions and Options.** For each script and vocal condition, we construct four questions following the Emotional Attitude Understanding (EAU), Mental State Attribution (MSA), Social Interaction Reasoning (SIR), and Recursive Pragmatic Reasoning (RPR) framework, each targeting a ToM ability dimension. Distractors represent plausible alternative mental-state interpretations rather than arbitrary choices. All questions undergo psychology experts review, with additional verification for shared scripts. Detailed protocols are provided in Appendix C.2. **Human Speech Recording.** We collect all human-recorded audio through a dedicated online platform, strictly following a standardized multi-stage protocol. Details are provided in Appendix C.3.

Evaluation Metrics We use Accuracy (Acc) as the metric and Exact Match over Four Questions (EM4) to measure consistency across 4 questions in each sample. For shared scripts with identical text but different vocal realizations, Pair Accuracy (PairAcc) and Flip Precision (FP) evaluate whether models distinguish the vocal conditions and adjust mental-state predictions. We report Vocal-Cue Gain, measuring the improvement from Text-only to Audio+Text, together with Fusion

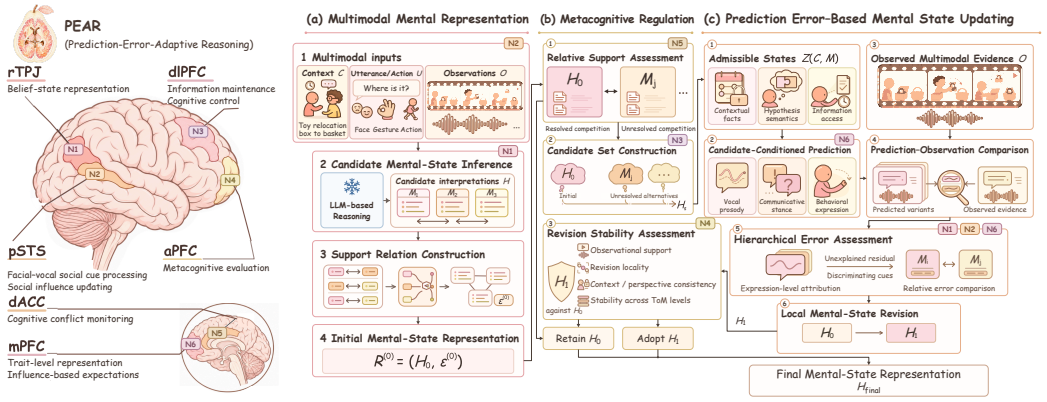

Figure 3: Overview of PEAR and its neurocognitive inspiration.

Gain (FG) and Complementarity Utilization (CU) to assess multimodal integration. FG compares A+T with the best unimodal setting, while CU measures how complementary information from text and audio is converted into performance gains. Full definitions are provided in Appendix C.4.

4 Methodology

PEAR is a training-free, brain-inspired ToM reasoning framework based on metacognitively regulated predictive coding. We formulate multimodal ToM reasoning as $\hat{M} = \mathcal{I}(C, U, O)$, where C denotes the social context, U the target utterance or key social behavior, O the observational evidence, and $\hat{M}$ the inferred mental state. In Overtone, O corresponds to vocal expression; in other ToM benchmarks, it may represent temporally organized behavioral evidence. Mental-state inference in the human brain can be viewed as an iterative predictive process: mental-state hypotheses generate top-down expectations about observable behavior, while bottom-up prediction errors drive revisions of the current interpretation. Metacognitive regulation further determines whether and how such revisions should be applied (Fernández Velasco & Loev, 2025; Koster-Hale & Saxe, 2013). Inspired by these mechanisms, PEAR comprises three modules, as shown in Figure 3.

4.1 Multimodal Mental Representation

PEAR jointly considers C, U, and O to compare candidate mental-state interpretations. Let $\mathcal{H} = \{M_1, \ldots, M_N\}$ denote the candidate mental-state set. Using a frozen multimodal backbone, PEAR constructs relative support among them and derives an initial representation:

$$E^{(0)} = \mathrm{Relate}_\theta(\mathcal{H} \mid C, U, O), \qquad H_0 = \mathcal{A}(\mathcal{H}, E^{(0)}), \qquad R^{(0)} = (H_0, E^{(0)}). \tag{1}$$

Here, H_0 denotes the initial mental-state judgment, while $E^{(0)}$ preserves support among competing interpretations, providing the basis for metacognitive regulation and prediction-driven updating.

4.2 Metacognitive Regulation

Mental-state reasoning requires regulating further processing when competing interpretations remain unresolved. Neurocognitive studies associate conflict monitoring with the dACC and metacognitive evaluation and control with the dlPFC and aPFC (Fleming & Dolan, 2014). Predictive-processing accounts further suggest that metacognition regulates whether inference should continue, terminate, or revise the current interpretation according to expected changes in prediction error (Fernández Velasco & Loev, 2025). Inspired by these mechanisms, PEAR uses metacognitive regulation to control perceptual processing and mental-state revision. Given $R^{(0)}$, where $M \parallel_0 H_0$ indicates that candidate M remains unresolved with respect to the initial interpretation H_0. PEAR identifies candidates insufficiently distinguished from H_0:

$$\mathcal{B}^{(0)} = \{M \in \mathcal{H} : M \neq H_0,\ M \parallel_0 H_0\}, \qquad \mathcal{H}_s = \{H_0\} \cup \mathcal{B}^{(0)}, \qquad g_{\mathrm{re}} = \mathbb{I}[\mathcal{B}^{(0)} \neq \emptyset]. \tag{2}$$

If $g_{\mathrm{re}} = 1$, PEAR initiates prediction over $\mathcal{H}_s$; otherwise, H_0 is retained. Metacognitive regulation provides two complementary controls: a front-end gate that determines whether and which candidates undergo further processing, and back-end that determines whether the revision is accepted.

4.3 Prediction Error-Based Mental State Updating

PEAR treats each $M \in \mathcal{H}_s$ as a predictive hypothesis, generates context-compatible observable outcomes, and compares them with the actual observation O:

$$\mathcal{O}_M = \bigcup_{z \in \mathcal{Z}(C,M)} P_\theta(U; M, z), \qquad M_i \succ_O M_j \iff \epsilon_\theta(M_i; C, U, O) \prec \epsilon_\theta(M_j; C, U, O). \quad (3)$$

Here, ϵ_θ denotes the residual prediction discrepancy and $\succ_O$ the resulting observational support. PEAR resolves discrepancies at the lowest explanatory level: expression-level differences are handled locally, while revisions to beliefs, intentions, or attitudes must preserve factual, temporal, and information-access constraints. A missing expected cue alone is insufficient to reject a hypothesis. The support relations produce a local revision $H_1 = \mathcal{U}_\theta(H_0, \mathcal{H}_s, \succ_O)$, which is accepted only after a metacognitive stability check, This separation between revision generation and acceptance prevents local observational discrepancies from producing unsupported high-level mental-state changes.

$$g_{\mathrm{st}} = \mathrm{Gate}_\theta(H_1, H_0 \mid C, U, O, \mathcal{Q}), \qquad H_{\mathrm{final}} = \begin{cases} H_1, & g_{\mathrm{re}} = 1 \wedge g_{\mathrm{st}} = 1, \\ H_0, & \text{otherwise.} \end{cases} \quad (4)$$

5 Experiments

5.1 Human Experiment

We recruit 48 participants for the human evaluation through a dedicated online platform to establish a reliable human baseline on Overtone. Full experimental details are provided in Appendix D.1.

5.2 Experiments 1 & 2: Overtone Evaluation and PEAR Assessment

Setting. (1) Evaluated Models. We evaluate 16 state-of-the-art models on Overtone. The model list and references are provided in Appendix D.2. For models with native reasoning capabilities, we use the strongest reasoning mode in their official implementations. **(2) PEAR Backbones.** PEAR is evaluated on four backbones: GPT-Audio-1.5 (OpenAI, 2026), Qwen3.5-Omni-Plus (Qwen Team, 2026), Gemini-3.1-Pro (Google DeepMind, 2026a), and Gemini-3.7-Flash (Google DeepMind, 2026b), under the same experimental configuration. **(3) Reasoning Methods.** For each backbone, we compare Direct, ToM CoT, PC ToM, and PEAR. Direct predicts from the full input, ToM CoT performs step-by-step mental-state reasoning, and PC ToM follows a prediction–comparison–error–update process. PEAR is our training-free framework. Prompts for ToM CoT and PC ToM are provided in Appendix L. **(4) Input Modalities.** Direct and ToM CoT are evaluated under Text-only, Audio-only, and Audio+Text settings, which differ only in the modality of the target utterance. PC ToM and PEAR use Audio+Text, as both reason over linguistic and vocal information.

Results and Analysis. (1) Overall Performance – Limited Mental-State Coherence and Vocal Grounding. Tables 2 and 3 report Audio+Text results on Overtone. OmniLLMs perform strongest overall. Across the 16 baseline models, however, mean accuracy is 64.86%, but EM4 falls to 29.57%, revealing substantial inconsistency across related mental-state questions. This limitation is more pronounced on shared scripts, where PairAcc and FP average only 5.61% and 28.23%, indicating that changes in vocal expression rarely lead to the corresponding mental-state distinctions. EAU is also the most challenging level (57.71%), highlighting difficulties with affective and pragmatic interpretation. Across finer-grained dimensions, baselines perform better on Mixed (72.31%) and Social (67.31%) than on Basic (62.24%) and Higher-order Pragmatic Attitudes (62.46%); Emotion is the weakest ToM ability (55.40%), followed by NLC (60.24%). PEAR substantially improves both reasoning performance and mental-state consistency, while achieving the highest accuracy across all four ToM levels and 11 reported dimensions. Detailed results are provided in the Appendix E.

Table 2: Audio+Text performance on Overtone across four ToM levels and Shared Pairs. **Bold** and underline indicate the best and second-best results. PEAR uses Gemini-3.7-Flash as its backbone.

Model	All Samples						Shared Pairs	
	EAU↑	MSA↑	SIR↑	RPR↑	Avg↑	EM4↑	PairAcc↑	Flip↑
Random Guess	24.25	26.28	26.46	23.39	25.09	0.23	6.18	8.09
Human	91.78	91.67	92.12	92.08	91.91	85.81	85.11	94.74
Speech Large Language Models (SpeechLLMs)								
GPT-Audio-1.5	59.05	71.02	72.11	74.66	69.21	35.92	3.06	29.68
Audio-Flamingo-Next	61.79	65.05	66.18	70.27	65.82	26.35	8.24	22.30
MOSS-Audio-8B	56.68	67.42	67.64	67.64	64.85	26.13	5.12	27.50
Step-Audio-R1.5	55.93	64.71	69.11	66.48	64.06	24.81	6.85	20.04
Kimi-Audio	56.42	59.01	63.25	61.71	60.10	17.27	14.30	28.51
MiMo-Audio	47.94	61.56	63.81	63.51	59.21	21.55	5.52	19.71
Audio-Reasoner	60.17	58.56	51.09	52.85	55.67	10.14	1.13	12.88
Omni Large Language Models (OmniLLMs)								
Gemini-3.7-Flash	76.13	78.68	80.44	81.98	79.31	56.34	9.84	50.00
Gemini-3.1-Pro	74.89	78.75	80.29	80.93	78.72	55.82	9.18	58.72
Qwen3.5-Omni-Plus	63.89	73.01	75.23	75.94	72.02	43.17	3.32	45.87
MiMo-V2.5	58.30	69.86	70.95	71.47	67.64	35.25	7.25	30.11
MiniCPM-o-4.5	54.88	67.57	68.92	70.91	65.57	30.56	1.60	29.27
Gemma	55.48	65.73	67.64	70.46	64.83	27.89	1.06	15.09
Nemotron-3-Nano-Omni	55.26	65.13	67.79	67.53	63.93	26.99	5.85	23.10
Human-Centric Omni Large Language Models (HC-Omni LLMs)								
HumanOmni V2	45.27	58.48	56.23	58.00	54.50	16.89	4.65	22.01
OmniSapiens 2.0	41.25	56.87	56.04	55.18	52.34	17.98	2.86	16.93
PEAR (Ours)	**86.56**	**87.73**	**88.44**	**88.10**	**87.71**	**82.51**	**23.34**	**67.50**

(2) Vocal-Cue Use and Multimodal Integration – Uneven Gains and Underused Complementarity. Figure 4 shows that adding audio improves average performance across all dimensions, with the largest gains on Emotion (+3.55 pp), NLC (+3.44 pp), and Basic Emotion (+2.92 pp). The gains vary markedly across models: Gemini-3.7-Flash improves across all dimensions, whereas Kimi-Audio gains 17.23 pp on NLC but drops 8.89 pp on Knowledge, and MiMo-Audio also regresses on several dimensions. Vocal cues provide useful information, but models differ in how reliably they exploit it. Figure 5 further examines this gap at the model level. As shown in panel (a), stronger performance does not necessarily imply larger vocal gains, and Audio+Text can even underperform Audio-only, as for Qwen3.5-Omni-Plus, indicating cross-modal interference. (b) shows a similar limitation in complementarity utilization: although the Oracle exceeds the best unimodal result for most models, only 8 of 16 achieve positive CU. These results suggest that the challenge is not merely extracting vocal cues, but integrating them reliably with linguistic evidence for ToM reasoning. Full results are in Appendix F. **(3) PEAR Improves ToM Reasoning through Staged Inference and Controlled Revision.** Table 4 compares PEAR with alternative reasoning strategies across four backbones. Explicit reasoning alone does not consistently improve performance: ToM CoT ranges from -1.60 to +2.98 pp, while PC ToM degrades four backbones by 4.41 pp on average. PEAR consistently improves all backbones and achieves more reliable mental-state revision, suggesting that staged inference and metacognitive regulation better exploit vocal variation while limiting harmful updates. Fine-grained results are provided in Appendix G.



Figure 4: Vocal gains across Affective–Pragmatic and ToM Ability Dimensions.

5.3 Experiment 3: Cross-Benchmark Evaluation

Setting. (1) Benchmarks. We evaluate PEAR on Hi-ToM (Wu et al., 2023), MuMA-ToM (Shi et al., 2025), and MMToM-QA (Jin et al., 2024) to examine its generalization beyond Overtone across text, higher-order, and multimodal ToM reasoning. PEAR retains the same core framework, with minor adaptations to the observation format. Details are provided in Appendix H. **(2) Baselines.**

Table 3: A+T performance across 4 Affective–Pragmatic Dimensions and 7 Ability Dimensions.

Model	Emotion Category				ToM Ability						
	Basic↑	Social↑	Mixed↑	Meta↑	Perc.↑	Know.↑	Bel.↑	Des.↑	Emo.↑	Int.↑	NLC↑
Random Guess	25.34	24.97	26.85	24.52	27.60	24.53	25.20	29.85	23.60	24.31	24.61
Human	92.88	91.69	93.15	91.05	91.83	91.86	92.96	92.91	91.70	91.76	90.90
Speech Large Language Models (SpeechLLMs)											
GPT-Audio-1.5	68.34	72.53	75.28	65.22	72.03	72.22	75.44	72.39	56.74	73.98	62.30
Audio-Flamingo-Next	62.54	68.28	77.96	62.53	65.47	69.46	67.71	67.04	61.24	67.45	63.38
MOSS-Audio-8B	60.43	66.66	77.13	62.88	65.59	66.21	67.71	70.52	53.87	70.29	59.50
Step-Audio-R1.5	59.34	66.44	75.46	62.07	62.25	66.08	64.97	69.40	53.25	70.19	60.42
Kimi-Audio	62.69	63.00	61.11	55.46	56.31	56.95	62.17	64.43	53.68	63.56	59.24
MiMo-Audio	54.86	60.72	68.98	58.17	58.17	65.96	63.03	64.05	44.76	66.60	51.64
Audio-Reasoner	52.90	60.16	59.72	52.61	53.34	57.70	57.07	56.72	58.86	54.46	53.10
Omni Large Language Models (OmniLLMs)											
Gemini-3.7-Flash	79.82	79.78	81.67	77.85	80.82	80.35	79.42	78.86	73.60	80.92	80.67
Gemini-3.1-Pro	79.14	79.69	82.78	76.40	80.94	80.73	78.61	77.86	72.10	81.16	79.21
Qwen3.5-Omni-Plus	70.22	72.75	78.89	70.67	70.05	75.59	75.50	76.00	61.24	76.85	67.80
MiMo-V2.5	64.19	70.34	73.70	66.02	68.19	69.09	72.27	72.14	56.30	71.79	63.44
MiniCPM-o-4.5	61.63	68.56	75.93	62.80	67.08	68.96	70.66	70.27	52.37	71.15	58.91
Gemma	63.22	69.94	68.52	60.51	66.83	66.96	67.49	66.04	53.56	68.75	63.38
Nemotron-3-Nano-Omni	59.34	65.91	75.19	62.23	63.37	66.58	68.57	69.15	51.62	69.33	58.21
Human-Centric Omni Large Language Models (HC-Omni LLMs)											
HumanOmni V2	49.66	57.22	63.80	52.90	56.31	63.83	56.69	59.45	43.95	61.33	43.67
OmniSapiens 2.0	47.59	54.97	60.83	50.99	53.84	62.45	55.78	58.58	39.20	60.89	39.04
PEAR (Ours)	**88.70**	**89.28**	**93.15**	**84.06**	**88.49**	**88.99**	**88.55**	**88.93**	**85.39**	**89.30**	**84.92**

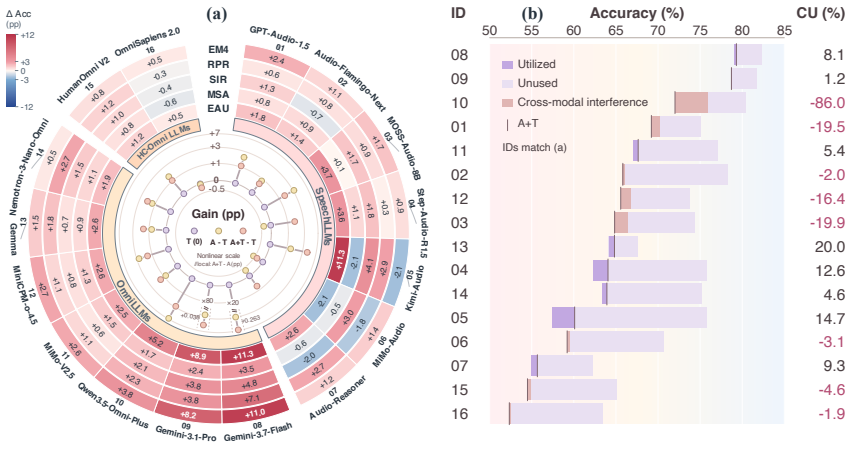

Figure 5: (a) Audio gain across ToM levels. (b) Cross-modal complementarity utilization.

We compare PEAR with SimToM (Wilf et al., 2024), BIP-ALM (Jin et al., 2024), LIMP (Shi et al., 2025), and AutoToM (Zhang et al., 2025), using GPT-4o (OpenAI, 2024) as the backbone. Following AutoToM, MuMA-ToM and MMToM-QA are converted into unified textual observations, while Hi-ToM retains its original text input, providing matched information across methods.

Results and Analysis. Generalizing from Vocal to Broader Social ToM Reasoning. Table 5 reports PEAR's performance across the three external ToM benchmarks. PEAR transfers effectively to settings that require integrating evolving social evidence, with particularly strong results on MMToM-QA and MuMA-ToM. On Hi-ToM, performance decreases as recursive depth increases, consistent with PEAR's focus on evidence-driven mental-state updating rather than explicit modeling of deeply nested perspectives. We therefore extend PEAR with perspective-taking, constraining each reasoning step to information available to the target agent. This substantially improves second and higher-order reasoning, showing that PEAR can support recursive ToM when combined with explicit perspective modeling. Full results and extension details are provided in Appendix I.

5.4 Experiment 4: Ablation Studies.

Setting. We ablate PEAR components on Gemini-3.7-Flash. **w/o Top-Down Prediction** removes predictions of observable expressions; **w/o Comparison** removes explicit prediction–observation comparison; **w/o Error-Guided Update** disables prediction error revision; and **w/o Metacognitive Regulation** retains the predictive-coding process but removes monitoring and control of candidate

Table 4: A+T results across backbones. Revision Precision (RP), the fraction of beneficial revisions.

	Direct	ToM CoT			PC ToM			PEAR		
Backbone	**Avg↑**	**Avg↑**	**ΔFP↑**	**RP↑**	**Avg↑**	**ΔFP↑**	**RP↑**	**Avg↑**	**ΔFP↑**	**RP↑**
GPT-Audio-1.5	69.21	67.61 (−1.60)	+13.71	37.76	62.18 (−7.03)	−7.72	23.16	**72.99** (+3.78)	**+17.60**	**49.86**
Qwen3.5-Omni-Plus	72.02	75.00 (+2.98)	+10.19	61.94	67.46 (−4.55)	−6.46	28.57	**79.37** (+7.36)	**+16.66**	**63.81**
Gemini-3.1-Pro	78.72	79.63 (+0.91)	+1.34	49.54	75.64 (−3.08)	−3.85	32.41	**86.28** (+7.56)	**+7.94**	**69.09**
Gemini-3.7-Flash	79.31	79.97 (+0.67)	+4.02	48.34	76.33 (−2.97)	−2.82	34.44	**87.71** (+8.40)	**+17.50**	**73.51**

Table 5: Cross-benchmark ToM results. Hi-ToM results are grouped into O0 (zero-order), O1 (first-order), and OH (higher-order). "–" denotes Perspective Taking is applied only to Hi-ToM.

Method	**Hi-ToM**	**MMToM-QA**	**MuMA-ToM**
SimToM	71.00 (O0 **100.00** O1 **77.50** OH 59.17)	51.00	47.63
BIP-ALM	14.50 (O0 10.00 O1 17.50 OH 15.00)	56.17	33.90
LIMP	6.50 (O0 5.00 O1 10.00 OH 5.83)	55.33	76.60
AutoToM	72.50 (O0 95.00 O1 75.00 OH 64.17)	83.00	81.44
PEAR	59.00 (O0 **100.00** O1 75.00 OH 40.00)	**83.50**	**91.11**
PEAR + Perspective Taking	**93.00** (O0 **100.00** O1 72.50 OH **97.50**)	–	–

updates. We also include a **Compute-Matched** variant with a comparable reasoning budget but no predictive-coding structure, with **Direct** as the reference.

Results and Analysis. Figure 6 and Table 6 show that full PEAR performs best when prediction, comparison, error-guided revision, and metacognitive regulation are combined. Removing metacognitive regulation produces more frequent revisions but lower RP and higher regression, showing that effective reasoning depends on selective rather than frequent updating. Revision Rate is the proportion of revised predictions. The Compute-Matched remains below full PEAR despite a comparable reasoning budget, indicating that PEAR's gains arise from structured mechanism interaction rather than additional computation. These results validate PEAR's mechanism synergy.

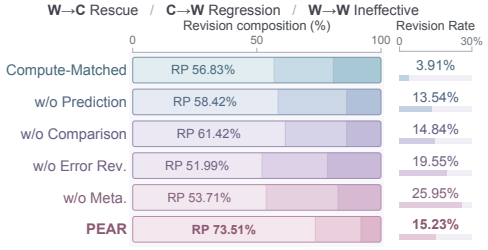

Figure 6: Revision statistics across ablations.

Table 6: Ablation study on PEAR components.

PEAR Components				**Performance**	
w/o TDP	w/o POC	w/o EGU	w/o MR	Avg↑	FP↑
–	✓	✓	✓	83.46	57.18
✓	–	✓	✓	84.76	60.09
✓	✓	–	✓	84.31	61.49
✓	✓	✓	–	85.74	60.69
Compute-Matched				80.59	50.14
✓	✓	✓	✓	**87.71**	**67.50**

6 Conclusion

In this work, we present **Overtone**, the first human-recorded multimodal ToM benchmark for studying how voice and pragmatics shape ToM reasoning. With a human baseline as reference, our evaluation of 16 state-of-the-art speech and omni-modal models reveals gaps in coherent ToM reasoning, vocal-cue utilization, and multimodal integration. To address these limitations, we develop **PEAR**, a training-free, brain-inspired reasoning framework that combines predictive coding with metacognitive regulation. Across diverse model backbones, PEAR consistently improves ToM reasoning and produces more reliable mental-state revisions. Its gains further transfer beyond Overtone to broader ToM benchmarks. Together, Overtone and PEAR establish a principled foundation for advancing multimodal social intelligence, shifting the frontier from recognizing observable social signals toward understanding the latent mental states that give human communication its meaning.

AI USE STATEMENT

In this work, we used OpenAI GPT-5.5 to assist with candidate script generation (Section 3). OpenAI GPT-5.6 assisted with paper translation and minor language polishing. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work involves human speech recording and human-subject evaluation. All speakers and participants were informed about the purpose of the study and compensated for their participation. Human-evaluation responses were collected anonymously, and participants were not exposed to unnecessary information beyond the assigned evaluation samples.

Overtone scenarios were created specifically for research purposes and do not contain private conversations or personally identifiable information. The mental-state represent controlled benchmark targets rather than definitive judgments about real individuals' internal states. Since models capable of inferring mental states from voice and multimodal behavior may raise privacy and misuse concerns, such systems should not be used for high-stakes decisions about individuals without appropriate human oversight and additional ethical safeguards.

REPRODUCIBILITY STATEMENT

We make extensive efforts to ensure the reproducibility of this work. For Overtone, we provide detailed descriptions of the benchmark construction process, including scenario design, script generation, human speech recording, quality control, question construction protocols. For PEAR, we describe the complete reasoning framework, including its modules, inference procedure, and implementation details. We further report all experimental settings, including evaluated models, backbones, baselines, prompts, evaluation metrics, and ablation protocols. Details of human evaluation procedures, data collection protocols, and additional analyses are provided in the Appendix. Upon acceptance, we will release the benchmark resources, code, and evaluation scripts to facilitate independent verification and future research.

REFERENCES

Sara B. Algoe and Jonathan Haidt. Witnessing excellence in action: The 'other-praising' emotions of elevation, gratitude, and admiration. *The Journal of Positive Psychology*, 4(2):105–127, 2009.

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1: 0064, 2017.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985.

Simon Baron-Cohen, Michelle O'Riordan, Valerie Stone, Rosie Jones, and Kate Plaisted. Recognition of faux pas by normally developing children and children with asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, 29(5):407–418, 1999.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, 10: 2905, 2020.

Geoffrey Bird and Essi Viding. The self to other model of empathy: Providing a new framework for understanding empathy impairments in psychopathy, autism, and alexithymia. *Neuroscience & Biobehavioral Reviews*, 47:520–532, 2014.

Brigitte Biró, Renáta Cserjési, Natália Kocsel, Attila Galambos, Kinga Gecse, Lilla Nóra Kovács, Dániel Baksa, Dóra Dobos, Gabriella Juhász, and Gyöngyi Kökönyei. Unveiling the role of theory of mind: Neural response to emotional stimuli in context. *Affective Science*, 6(2):340–355, 2025.

Diane Blakemore. Echo questions: A pragmatic account. *Lingua*, 94(4):197–211, 1994.

Sandra Bosacki and Janet Wilde Astington. Theory of mind in preadolescence: Relations between social understanding and social competence. *Social Development*, 8(2):237–255, 1999.

Stephanie Burnett, Geoffrey Bird, Jorge Moll, Chris Frith, and Sarah-Jayne Blakemore. Development during adolescence of the neural processing of social emotion. *Journal of Cognitive Neuroscience*, 21(9):1736–1750, 2009.

Jingyi Chen, Zhimeng Guo, Jiyun Chun, Pichao Wang, Andrew Perrault, and Micha Elsner. Do audio LLMs really LISTEN, or just transcribe? measuring lexical vs. acoustic emotion cues reliance. In *Proc. Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 5848–5877, 2026.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. ToMBench: Benchmarking theory of mind in large language models. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15959–15983, 2024.

Coralie Chevallier, Ira Noveck, Francesca Happé, and Deirdre Wilson. What's in a voice? prosody as a test case for the theory of mind account of autism. *Neuropsychologia*, 49(3):507–517, 2011.

Rhiannon Corcoran, Gavin Mercer, and Christopher D. Frith. Schizophrenia, symptomatology and social inference: Investigating "theory of mind" in people with schizophrenia. *Schizophrenia Research*, 17(1):5–13, 1995.

Pedro Corrêa, João Lima, Victor Moreno, Lucas H. Ueda, and Paula Dornhofer Paro Costa. Evaluating emotion recognition in spoken language models on emotionally incongruent speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17952–17956, 2026.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao.

MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *CoRR*, abs/2604.27393, 2026.

Murray J. Dyck, Kara Ferguson, and Ian M. Shochet. Do autism spectrum disorders differ from each other and from non-spectrum disorders on emotion recognition tests? *European Child & Adolescent Psychiatry*, 10(2):105–116, 2001.

Pablo Fernández Velasco and Slawa Loev. Metacognitive feelings: A predictive-processing perspective. *Perspectives on Psychological Science*, 20(4):691–713, 2025.

Stephen M. Fleming and Raymond J. Dolan. The neural basis of metacognitive ability. In Stephen M. Fleming and Christopher D. Frith (eds.), *The Cognitive Neuroscience of Metacognition*, pp. 245–265. Springer, 2014.

Xiaoxue Gao, Eshin Jolly, Hongbo Yu, Huiying Liu, Xiaolin Zhou, and Luke J. Chang. The psychological, computational, and neural foundations of indebtedness. *Nature Communications*, 15 (1):68, 2024.

Gemma Team. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, Chao-Han Huck Yang, Sungwon Kim, Ramani Duraiswami, Dinesh Manocha, Mohammad Shoeybi, Bryan Catanzaro, Ming-Yu Liu, and Wei Ping. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *CoRR*, abs/2604.10905, 2026.

Noah D. Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013.

Google DeepMind. Gemini 3.1 Pro model card. Google DeepMind, 2026a. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`. Published February 19, 2026; accessed September 19, 2026.

Google DeepMind. Gemini 3.7 Flash model card. Google DeepMind, 2026b. URL `https://deepmind.google/models/model-cards/gemini-3-7-flash/`. Published August 13, 2026; accessed September 19, 2026.

Francesca G. E. Happé. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2):129–154, 1994.

Paul L. Harris. *Children and Emotion: The Development of Psychological Understanding*. Basil Blackwell, 1989.

Paul L. Harris, Kara Donnelly, Gabrielle R. Guz, and Rosemary Pitt-Watson. Children's understanding of the distinction between real and apparent emotion. *Child Development*, 57(4):895–909, 1986.

Paul L. Harris, Carl N. Johnson, Deborah Hutton, Giles Andrews, and Tim Cooke. Young children's theory of mind and emotion. *Cognition and Emotion*, 3(4):379–400, 1989.

Erin A. Heerey, Dacher Keltner, and Lisa M. Capps. Making sense of self-conscious emotion: Linking theory of mind and emotion in children with autism. *Emotion*, 3(4):394–400, 2003.

Laurence R. Horn. Metalinguistic negation and pragmatic ambiguity. *Language*, 61(1):121–174, 1985.

Sean Dae Houlihan, Max Kleiman-Weiner, Luke B. Hewitt, Joshua B. Tenenbaum, and Rebecca Saxe. Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251): 20220047, 2023.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16077–16102, 2024.

Melanie Killen, Kelly Lynn Mulvey, Cameron Richardson, Noah Jampol, and Amanda Woodward. The accidental transgressor: Morally-relevant theory of mind. *Cognition*, 119(2):197–215, 2011.

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B. Tenenbaum, and Yejin Choi. Hypothesis-driven theory-of-mind reasoning for large language models. In *Proc. Conference on Language Modeling (COLM)*, 2025.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-Audio technical report. *CoRR*, abs/2504.18425, 2025.

Anna M. Kołodziejczyk and Sandra L. Bosacki. Young-school-aged children's use of direct and indirect persuasion: Role of intentionality understanding. *Psychology of Language and Communication*, 20(3):292–315, 2016.

Jorie Koster-Hale and Rebecca Saxe. Theory of mind: A neural prediction problem. *Neuron*, 79(5): 836–848, 2013.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5872–5877, 2019.

Zhu Li, Yuqing Zhang, Xiyuan Gao, Shekhar Nayak, and Matt Coler. Leveraging large language models for sarcastic speech annotation in sarcasm detection. In *Proc. Interspeech*, pp. 3973–3977, 2025.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8679–8696, 2025.

Shicheng Ma, Wenqian Cui, and Irwin King. SpeechSense: A paralinguistic-focused dataset for fine-grained speech sentiment analysis. In *Proc. ACM International Conference on Multimedia (ACM MM)*, 2026.

Andrew N. Meltzoff. Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31(5):838–850, 1995.

Wulf-Uwe Meyer, Rainer Reisenzein, and Achim Schützwohl. Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21(3):251–274, 1997.

Eun-Ju Noh. Echo questions: Metarepresentation and pragmatic enrichment. *Linguistics and Philosophy*, 21(6):603–628, 1998.

Greg J. Norman, Elizabeth Necka, and Gary G. Berntson. The psychophysiology of emotions. In *Emotion Measurement*, pp. 83–98. Elsevier, 2016.

NVIDIA. Nemotron 3 Nano Omni: Efficient and open multimodal intelligence. *CoRR*, abs/2604.24954, 2026.

Keane Ong, Sabri Boughorbel, Luwei Xiao, Chanakya Ekbote, Wei Dai, Ao Qu, Jingyao Wu, Rui Mao, Ehsan Hoque, Erik Cambria, Gianmarco Mengaldo, and Paul Pu Liang. OmniSapiens: A foundation model for social behavior processing via heterogeneity-aware relative policy optimization. In *Proc. International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

OpenAI. GPT-4o system card. OpenAI system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenAI. GPT-Audio-1.5. OpenAI API model documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-audio-1.5. Accessed September 19, 2026.

Josef Perner and Heinz Wimmer. "john thinks that mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3):437–471, 1985.

Ann T. Phillips, Henry M. Wellman, and Elizabeth S. Spelke. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition*, 85(1):53–78, 2002.

Bradford H. Pillow. Early understanding of perception as a source of knowledge. *Journal of Experimental Child Psychology*, 47(1):116–129, 1989.

Francisco Pons, Paul L. Harris, and Marc de Rosnay. Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European Journal of Developmental Psychology*, 1(2):127–152, 2004.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

Livia Qian and Gabriel Skantze. Aligning backchannel and dialogue context representations via contrastive LLM fine-tuning. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13817–13833, 2026.

Qwen Team. Qwen3.5-Omni technical report. *CoRR*, abs/2604.15804, 2026.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *Proc. International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227, 2018.

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

Betty M. Repacholi and Alison Gopnik. Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, 33(1):12–21, 1997.

James A. Russell. Emotion, core affect, and psychological construction. *Cognition and Emotion*, 23 (7):1259–1283, 2009.

Klaus R. Scherer, Johan Sundberg, Lucas Tamarit, and Gláucia L. Salomão. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, 29 (1):218–235, 2015.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13960–13980, 2023.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. MuMA-ToM: Multi-modal multi-agent theory of mind. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 1510–1519, 2025.

Craig A. Smith and Phoebe C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838, 1985.

Dan Sperber and Deirdre Wilson. Irony and the use-mention distinction. In Peter Cole (ed.), *Radical Pragmatics*, pp. 295–318. Academic Press, 1981.

June Price Tangney, Rowland S. Miller, Laura Flicker, and Deborah Hill Barlow. Are shame, guilt, and embarrassment distinct emotions? *Journal of Personality and Social Psychology*, 70(6): 1256–1269, 1996.

Csilla Tatár, Jonathan R Brennan, Jelena Krivokapić, and Ezra Keshet. Does prosody mark sarcasm early in an utterance? a production and perception study, including listeners who self-identified as being on the autism spectrum. *Journal of the International Phonetic Association*, 56:e1, 2026.

Jessica L. Tracy and Richard W. Robins. Putting the self into self-conscious emotions: A theoretical model. *Psychological Inquiry*, 15(2):103–125, 2004.

Jessica L. Tracy and Richard W. Robins. The psychological structure of pride: A tale of two facets. *Journal of Personality and Social Psychology*, 92(3):506–525, 2007.

Emilio Villa-Cueva, S M Masrur Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. MoMentS: A comprehensive multimodal benchmark for theory of mind. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*, pp. 22591–22611, 2025.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

Qiongqiong Wang, Hardik Bhupendra Sailor, Tianchi Liu, Wenyu Zhang, Muhammad Huzaifah, Nattadaporn Lertcheva, Shuo Sun, Nancy F. Chen, Jinyang Wu, and AiTi Aw. Benchmarking contextual and paralinguistic reasoning in speech-LLMs: A case study with in-the-wild data. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*, pp. 14133–14148, 2025.

Nigel G. Ward and Divette Marco. A collection of pragmatic-similarity judgments over spoken dialog utterances. In *Proc. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 154–163, 2024.

Philip Watkins, Jason Scheer, Melinda Ovnicek, and Russell Kolts. The debt of gratitude: Dissociating gratitude and indebtedness. *Cognition and Emotion*, 20(2):217–241, 2006.

Henry M. Wellman and David Liu. Scaling of theory-of-mind tasks. *Child Development*, 75(2): 523–541, 2004.

Henry M. Wellman and Jacqueline D. Woolley. From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3):245–275, 1990.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8292–8308, 2024.

Deirdre Wilson. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743, 2006.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983.

Ellen Winner and Sue Leekam. Distinguishing irony from deception: Understanding the speaker's second-order intention. *British Journal of Developmental Psychology*, 9(2):257–270, 1991.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Proc. Findings of the Association for Computational Linguistics: EMNLP*, pp. 10691–10706, 2023.

Xiaomi LLM-Core Team. MiMo-Audio: Audio language models are few-shot learners. *CoRR*, abs/2512.23808, 2025.

Xiaomi MiMo. MiMo-V2.5. Xiaomi MiMo model documentation, 2026. URL `https://mimo.mi.com/models/en-US/mimo-v2.5`. Accessed September 19, 2026.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-Reasoner: Improving reasoning capability in large audio language models. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 23829–23851, 2025.

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Donghua Yu, Jun Zhan, Kang Yu, Kexin Huang, Liwei Fan, Mingshu Chen, Qinyuan Cheng, Ruixiao Li, Shimin Li, Songlin Wang, Xingjian Zhao, Yang Gao, Yitian Gong, Yiyang Zhang, Zhe Xu, and Xipeng Qiu. MOSS-Audio technical report. *CoRR*, abs/2606.01802, 2026.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. HumanOmniV2: From understanding to omni-modal reasoning with context. *CoRR*, abs/2506.21277, 2025.

Iddo Yosha, Gallil Maimon, and Yossi Adi. StressTest: Can YOUR speech LM handle the stress? In *Proc. Findings of the Association for Computational Linguistics: ACL*, pp. 1264–1285, 2026.

Yuxin Zhang, Xiangyu Tony Zhang, Daijiao Liu, Fei Tian, Yayue Deng, Jun Chen, Qingjian Lin, Haoyang Zhang, Yuxin Li, Jinglan Gong, Yechang Huang, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Gang Yu, Xiangyu Zhang, and Daxin Jiang. Step-Audio-R1.5 technical report. *CoRR*, abs/2604.25719, 2026.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, Shunchi Zhang, and Tianmin Shu. AutoToM: Scaling model-based mental inference via automated agent modeling. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Li Zhou, Lutong Yu, You Lyu, Yihang Lin, Zefeng Zhao, Junyi Ao, Yuhao Zhang, Benyou Wang, and Haizhou Li. EchoMind: An interrelated multi-level benchmark for evaluating empathetic speech language models. In *Proc. International Conference on Learning Representations (ICLR)*, 2026.

Appendix

A Dataset Statistics **18**
B Design Dimensions of Overtone **18**
B.1 Affective & Pragmatic Dimension 18
B.2 8 Theory-of-Mind Tasks 18
B.3 Theory-of-Mind Ability Dimension 22
C Benchmark Construction and Evaluation **24**
C.1 Expert-Guided Script Generation and Quality Control 24
C.2 Question and Option Construction 27
C.3 Human Speech Recording 31
C.4 Evaluation Metrics 33
D Additional Experimental Details **36**
D.1 Human Experiment 36
D.2 Experimental Setup 37
E Fine-Grained Results on Overtone **38**
E.1 Task Dimensions 39
E.2 Affective and Pragmatic Dimensions 39
E.3 Theory-of-Mind Subabilities 40
F Audio-Only and Text-Only Results **41**
F.1 Audio-Only Results 41
F.2 Text-Only Results 42
F.3 Audio+Text versus Text-Only: Fine-Grained Comparisons 42
G Complete Results of ToM CoT and PC-ToM **46**
H PEAR Adaptation Across Diverse Theory-of-Mind Benchmarks **47**
H.1 Unified Predictive-Coding Formulation 47
H.2 MuMA-ToM: Hybrid Predictive Reasoning Adapter 48
H.3 MMToM-QA: Prediction-Error Overlay 48
H.4 Hi-ToM: Perspective-Aware Adapter 49
I Additional Experimental Results **49**
I.1 Hi-ToM Full Results 49
I.2 MuMA-ToM Full Results 49
I.3 MMToM-QA Full Results 50
J Dataset Examples **51**
K Shared-Script Pair Examples **63**
L Prompts **66**
L.1 Direct 66
L.2 ToM CoT 67
L.3 PC ToM 69

A Dataset Statistics

Table 7: Dataset composition by affective-pragmatic dimensions and ToM abilities.

Affective-pragmatic dimensions	
Dimension	**Questions**
Basic Emotion	**2,656**
Happy (452); Sad (440); Angry (452)	
Bad (444); Surprised (448); Fearful (420)	
Social Emotion	**3,200**
Guilty (804); Shame (784)	
Embarrassed (808); Pride (804)	
Mixed Social Emotion	**1,080**
Indebtedness (1,080)	
Higher-order Pragmatic Attitudes	**3,720**
Admiration (572); Irony (872)	
Sincere (572); Opposite (568)	
Echoic question (292); Endorsing attitude (288)	
Metalinguistic negation (280); Negative (276)	
Total	**10,656**

ToM abilities	
Category	**Questions**
Percepts	808
Knowledge	799
Beliefs	1,861
Desires	804
Emotions	1,602
Intentions	2,925
Non-Literal Communication	1,857
Total	**10,656**

B Design Dimensions of Overtone

B.1 Affective & Pragmatic Dimension

Overtone organizes affective and pragmatic expressions into four dimensions: *Basic Emotion*, *Social Emotion*, *Higher-order Pragmatic Attitudes*, and *Mixed Social Emotion*, comprising 19 fine-grained labels in total. These dimensions characterize not only the affective states conveyed by speakers, but also the social evaluations and communicative attitudes through which their utterances are interpreted. The taxonomy serves as an operational organization of the benchmark rather than assuming that all labels correspond to discrete emotions or involve an identical level of mental-state reasoning. Tables 8–11 summarize the psychological, cognitive, and pragmatic basis of each label.

B.2 8 Theory-of-Mind Tasks

Overtone adopts the eight task categories used in ToMBench (Chen et al., 2024) and grounds scenario design in their underlying psychological literature. Task categories characterize the social-cognitive structure of each scenario, whereas individual questions target specific mental-state reasoning abilities. These levels are complementary rather than in one-to-one correspondence. Below, we describe the psychological basis, characteristic scenario structure, and inferential demands of each task.

Unexpected Outcome Test. The Unexpected Outcome Test examines whether an observer can explain an emotional response that departs from what a situation would ordinarily suggest. Dyck et al. (2001) employed this paradigm to assess explanations of unexpected emotion-related behavior in short narratives. Its central premise is that the apparent desirability of an event does not uniquely determine its psychological significance for the person experiencing it: an ostensibly favorable outcome may elicit disappointment, whereas an apparently adverse change may bring relief. Understanding such reactions requires moving beyond familiar situation–emotion associations to consider the character's desires, beliefs, expectations, or unexpressed concerns. The task therefore emphasizes explaining why an event has a particular emotional significance for an individual, rather than identifying an emotion category alone. A successful interpretation connects the observed response

Table 8: **Basic Emotion.** Psychological characterization of the basic-emotion labels in Overtone.

Label	Psychological Characterization	References
happy	Positive affect associated with an outcome, event, or interpersonal state appraised as desirable or consistent with the individual's goals and concerns. Its interpretation therefore depends on whose goals or expectations are satisfied rather than on the objective valence of the event alone.	(Smith & Ellsworth, 1985; Harris, 1989)
sad	Negative affect associated with perceived loss, an unfavorable outcome, or failure to attain a valued goal. Understanding sadness requires relating the event to the individual's own expectations, desires, and concerns rather than simply identifying an objectively negative situation.	(Smith & Ellsworth, 1985; Harris, 1989)
angry	Negative affect associated with perceived wrongdoing, obstruction, or responsibility for an undesirable outcome. Its interpretation depends partly on how the individual attributes agency, control, and blame for what has occurred.	(Smith & Ellsworth, 1985)
bad	A broadly unpleasant affective state that is not necessarily differentiated into a more specific emotion such as sadness, anger, or fear. The label captures negative experiential valence rather than propositional negation or a moral judgment about an event.	(Russell, 2009)
surprised	An affective response elicited when an event violates or substantially departs from an existing expectation. Its interpretation therefore requires reconstructing what the individual previously expected and how the observed event departs from that expectation.	(Meyer et al., 1997)
fearful	A negative state associated with perceived threat, uncertainty, or anticipated undesirable consequences. Fear attribution depends on what the individual believes may happen and on their perceived ability to avoid or cope with that outcome.	(Smith & Ellsworth, 1985)

Table 9: **Social Emotion.** Psychological characterization of the social-emotion labels in Overtone.

Label	Psychological Characterization	References
guilty	A self-conscious negative emotion centered primarily on a specific action and its consequences. Guilt commonly involves perceived responsibility for another person's harm or disadvantage and may motivate apology, compensation, or interpersonal repair.	(Tangney et al., 1996; Tracy & Robins, 2004)
shame	A self-conscious negative emotion in which an undesirable event is interpreted as reflecting negatively on the self or an important aspect of one's identity. Unlike guilt, the evaluative focus extends beyond a particular behavior to a broader representation of the self.	(Tangney et al., 1996; Tracy & Robins, 2004)
embarrassed	An aversive social emotion arising when one's behavior, performance, or social presentation conflicts with perceived social expectations or attracts unwanted attention. Its interpretation often requires representing how the individual believes they are being perceived by others.	(Tangney et al., 1996; Heerey et al., 2003)
pride	A positive self-conscious emotion arising when an individual attributes a valued achievement or socially desirable outcome to their own actions, abilities, or characteristics. It therefore involves both evaluation of the outcome and attribution of personal contribution.	(Tracy & Robins, 2004; 2007)

to a plausible mental state while remaining consistent with the available context; inventing arbitrary background information to resolve the apparent discrepancy does not constitute a well-supported explanation.

Scalar Implicature Task. The Scalar Implicature Task examines how listeners infer unstated meanings by reasoning about a speaker's choice among expressions that differ in informativeness.

Table 10: **Higher-order Pragmatic Attitudes.** Psychological and pragmatic characterization of the higher-order pragmatic-attitude labels in Overtone.

Label	Psychological / Pragmatic Characterization	References
admiration	A positive other-directed evaluation elicited by another person's excellence, ability, achievement, or exemplary behavior. Unlike generic positive affect, admiration is directed toward the perceived qualities or accomplishments of another individual.	(Algoe & Haidt, 2009)
irony	A nonliteral communicative attitude in which the speaker evokes or echoes a thought, expectation, or prior expression while conveying dissociation from it. Interpretation therefore requires distinguishing the literal content from the evaluative stance the speaker intends to communicate.	(Sperber & Wilson, 1981; Wilson, 2006)
sincere	An attitude in which the stance overtly expressed by the speaker is aligned with the speaker's actual belief, feeling, or evaluation. Sincerity concerns correspondence between expression and internal state rather than factual correctness, politeness, or positive affect.	(Chevallier et al., 2011)
opposite	A communicative state in which the attitude conveyed at the surface level contrasts with the speaker's underlying affective or evaluative state. Such a discrepancy may serve concealment, reassurance, politeness, or conflict avoidance and therefore need not function in the same way as irony.	(Chevallier et al., 2011; Winner & Leekam, 1991)
echoic question	An interrogative repetition or reformulation of previously expressed content. Beyond requesting information, an echoic question can communicate surprise, doubt, disbelief, or challenge toward the representation being repeated.	(Blakemore, 1994; Noh, 1998)
endorsing attitude	An attitude in which the speaker adopts or supports a previously expressed or contextually available thought as compatible with their own position. It contrasts with echoic uses in which the speaker distances themselves from or rejects the represented content.	(Sperber & Wilson, 1981; Wilson, 2006)
metalinguistic negation	Negation directed primarily at the appropriateness of a linguistic expression rather than simply at the truth of the proposition it expresses. The rejected aspect may concern wording, degree, implication, register, or another property of how the proposition has been formulated.	(Horn, 1985)
negative	Propositional rejection or denial of a contextually relevant claim. In contrast to metalinguistic negation, the primary target is the proposition under discussion rather than the appropriateness of the linguistic form used to express it.	(Horn, 1985)

Table 11: **Mixed Social Emotion.** Psychological characterization of the mixed social-emotion label in Overtone.

Label	Psychological Characterization	References
indebtedness	A relational state arising after receiving another person's help, benefit, or sacrifice, in which appreciation of the benefit can coexist with guilt, obligation, or perceived pressure to reciprocate. Unlike gratitude alone, indebtedness is particularly sensitive to the perceived cost incurred by the benefactor and to expectations of future repayment. Its interpretation therefore requires reasoning about both the received benefit and the interpersonal obligations that the recipient believes the interaction has created.	(Watkins et al., 2006; Gao et al., 2024)

For example, although *some* is semantically compatible with *all*, choosing the weaker expression can, in an appropriate context, support the inference *some but not all*. This inference is defeasible rather than a fixed lexical rule. Goodman & Stuhlmüller (2013) showed that manipulating a speaker's access to information changes listeners' interpretations: when the speaker has only

partial knowledge, an otherwise expected scalar implicature may weaken or disappear. The task consequently requires integrating the available linguistic alternatives, the speaker's knowledge, and the communicative goal to determine whether a weaker statement conveys rejection of a stronger proposition or merely reflects uncertainty about it. Its central demand is to infer the speaker's epistemic position from their linguistic choices, rather than automatically equating a less informative expression with the negation of a more informative one.

Persuasion Story Task. The Persuasion Story Task concerns attempts to influence another person's decisions or behavior through communication, with an emphasis on how persuasive strategies reflect an understanding of the addressee's mental states. Kołodziejczyk & Bosacki (2016) investigated the relationship between children's understanding of intentionality and their use of direct and indirect persuasive strategies in narratives involving parents and peers. Such scenarios typically juxtapose the persuader's desired outcome with an addressee who has not agreed, holds different preferences, or has unresolved concerns. Interpreting the interaction requires distinguishing the two perspectives, identifying the change sought by the persuader, and explaining why a particular appeal might affect the addressee's judgment. Effective other-oriented persuasion goes beyond reiterating one's own wishes: it takes into account what the other person values, already believes, and may accept within the relationship. The task thus concerns not only what a character wants, but also how their representation of another person's mind informs the communicative means chosen to pursue that goal.

False Belief Task. The False Belief Task examines whether an observer can distinguish a character's representation of the world from its actual state and use that representation to explain or predict behavior. In the classic paradigm of Wimmer & Perner (1983), a character retains an earlier belief because they do not witness a subsequent change. Although the observer knows the current state of affairs, they must predict the character's action from the information available to that character. Successful reasoning therefore requires tracking both event order and person-specific information access, without attributing the observer's knowledge to every participant. Perner & Wimmer (1985) extended this structure to second-order beliefs, in which one person represents what another person believes. Even when both characters know a fact, one may incorrectly believe that the other remains uninformed. The task therefore requires maintaining distinct mental representations and, in higher-order cases, separating *what is true*, *what a person believes*, and *what that person believes someone else believes*.

Ambiguous Story Task. The Ambiguous Story Task examines how an observer constructs a justified mental-state interpretation when a social narrative leaves the causes of a character's behavior underspecified. Bosacki & Wilde Astington (1999) studied the understanding of thoughts and emotions in social situations and its relationship with social competence in preadolescence. In such scenarios, the same action may be compatible with different motives, feelings, or interpersonal attitudes, so an isolated behavior or utterance rarely determines a unique interpretation. Reasoning requires integrating the interaction history, the characters' relationship, and surrounding actions to assess which explanation best accounts for the available information. Here, ambiguity concerns uncertain social meaning and psychological causation, rather than lexical or syntactic ambiguity alone. Nor does it imply that all interpretations are equally plausible. The central demand is to make context-sensitive mental-state attributions under incomplete information while distinguishing supported inferences from assumptions about beliefs, emotions, or intentions that the narrative does not establish.

Hinting Test. The Hinting Test examines whether an observer can recover a speaker's communicative intention from an indirect utterance. Corcoran et al. (1995) introduced a task assessing the ability to infer the intentions underlying indirect speech. The central question is what a speaker wants the listener to understand or do beyond the literal content of the utterance. A statement about a situation, a description of one's condition, or an apparent request for information may also function as a request, reminder, or suggestion. Recognizing this function requires relating the utterance to the speaker's current needs, the listener's capacity to respond, and the context shared by both participants. The task is therefore not simply to assign an additional hidden meaning to a sentence, but to explain why that sentence is addressed to this listener at this moment. In contrast to persua-

sion, which emphasizes strategies for changing another person's position, hinting primarily concerns recognizing a communicative goal that remains unstated.

Strange Story Task. The Strange Story Task examines the interpretation of nonliteral language and its psychological motivations in complex social situations. Happé (1994) introduced naturalistic narratives that ask why a character makes a particular utterance, encompassing phenomena such as deception, white lies, jokes, pretense, and irony. The difficulty lies in the potential divergence among what is literally said, what the speaker actually believes, and the social effect they intend to achieve. A statement inconsistent with reality may reflect misunderstanding, deliberate deception, an attempt to protect another person's feelings, or irony. Distinguishing these possibilities requires considering the speaker's knowledge and intentions, together with their expectations about the listener's interpretation. Merely judging whether an utterance is true or detecting negative wording is therefore insufficient. The task instead requires a context-appropriate mental-state explanation of why an expression serves a particular function within the ongoing interaction. Its scope extends across multiple forms of complex communication rather than reducing to irony recognition alone.

Faux Pas Recognition Test. The Faux Pas Recognition Test examines whether an observer can identify an inadvertently inappropriate remark and explain its interpersonal consequences. Baron-Cohen et al. (1999) used this paradigm to assess advanced social understanding, requiring sensitivity to both the speaker's limited knowledge and the listener's emotional perspective. Recognizing a faux pas entails more than noticing that a remark is socially inappropriate: the observer must determine what the speaker does not know, why they fail to anticipate the impact of their words, and why a particular listener may feel embarrassed or hurt. An utterance may be factually accurate and contain no explicit insult, yet still have negative consequences because of the listener's experiences or personal investment. Crucially, these consequences must be distinguished from the speaker's intention; causing offense does not itself establish an intention to offend. The task thus jointly engages knowledge attribution, intention understanding, and affective perspective-taking, rather than general judgments of politeness or speech fluency.

B.3 Theory-of-Mind Ability Dimension

Different from the task dimension, which characterizes the social-cognitive structure of a scenario, the ability dimension specifies the mental-state competence required to answer each question. We build this taxonomy upon the Abilities in Theory-of-Mind Space (ATOMS) framework (Beaudoin et al., 2020), which systematically organizes Theory-of-Mind (ToM) abilities across multiple forms of mental-state understanding. Following this framework, Overtone organizes its question-level abilities into seven major dimensions: *Emotion*, *Desire*, *Intention*, *Percept*, *Knowledge*, *Belief*, and *Non-literal*. Considering the socially situated and speech-based setting of Overtone, we select 33 fine-grained sub-abilities that can be meaningfully instantiated through social context, vocal expression, and communicative behavior. Detailed descriptions of the seven dimensions are provided in Table 13 (Emotion), Table 14 (Desire), Table 15 (Intention), Table 16 (Percept), Table 17 (Knowledge), Table 18 (Belief), and Table 19 (Non-literal).

Table 12 provides an overview of the seven ToM ability dimensions, their corresponding sub-abilities, and their cognitive focus. Detailed psychological descriptions and references for each sub-ability are provided in the following tables.

Each question in Overtone is assigned to one primary sub-ability according to the central mental-state operation required for solving it. This assignment does not imply that these abilities operate independently. A single social inference may involve several mental states, such as using another person's beliefs to predict their emotion, or combining perceptual access with communicative intention. The annotation instead identifies the principal inferential demand of each question. Notably, *Emotion* here denotes a question-level ToM ability and is distinct from the audio-level affective and pragmatic taxonomy described in Appendix B.1.

A1 Emotion. Emotion-related ToM concerns reasoning about another person's affective state in relation to their beliefs, desires, evaluations, and social circumstances rather than merely recognizing an observable emotional expression. Developmental studies show that emotion understanding progressively incorporates situational causes, desires, beliefs, distinctions between felt and displayed

Table 12: **Overview of ToM ability dimensions in Overtone.** The taxonomy contains seven major dimensions and 33 fine-grained sub-abilities grounded in psychological ToM literature.

Dimension	Cognitive Focus	#Sub-abilities
Emotion	Reasoning about others' affective states, including emotional causes, discrepant emotions, hidden emotions, and emotion regulation.	7
Desire	Understanding others' preferences, goals, and desire-dependent behavior.	4
Intention	Inferring goal-directed actions and communicative intentions.	4
Percept	Representing others' perceptual access and perspective.	2
Knowledge	Reasoning about others' information access and epistemic states.	4
Belief	Representing false beliefs, higher-order beliefs, and belief-dependent behavior.	6
Non-literal	Understanding nonliteral meanings, deception, humor, and social appropriateness.	6

emotion, mixed emotions, moral evaluations, and emotion regulation (Harris, 1989). Overtone distinguishes seven sub-abilities within this dimension.

A2 Desire. Desire understanding concerns representing what another person wants or prefers and using that representation to explain their emotions and behavior. Developmental studies demonstrate that children can distinguish another person's desire from their own and use desires to interpret goal-directed behavior before they consistently succeed on more demanding false-belief tasks (Wellman & Woolley, 1990). Overtone distinguishes four sub-abilities.

A3 Intention. Intention understanding concerns the outcomes an agent is trying to bring about through action. A fundamental distinction is that an intended goal cannot always be recovered directly from the observable outcome: an unsuccessful action may still reveal what the actor intended, whereas similar behaviors may arise from different intentions (Meltzoff, 1995; Killen et al., 2011). Overtone distinguishes four sub-abilities.

A4 Percept. Percept-related ToM requires representing the social situation from another person's perceptual access rather than from the observer's own perspective. What a person can see or hear constrains both the information available to them and the actions they may subsequently take. Because Overtone centers on spoken interaction(Pillow, 1989), auditory perspective-taking is particularly relevant to this dimension.

A5 Knowledge. Knowledge-related ToM concerns what information another person has access to and how perceptual or communicative access establishes or limits that person's epistemic state. Developmental research shows that understanding the causal relation between perceptual access and knowledge is a fundamental component of social cognition (Pillow, 1989). Overtone distinguishes four sub-abilities.

A6 Belief. Belief reasoning requires representing another person's representation of the world even when it differs from reality or from one's own knowledge. Classic false-belief paradigms establish this capacity as a central component of explicit ToM (Wimmer & Perner, 1983), whereas higher-order tasks additionally require representing one person's belief about another person's mental state (Perner & Wimmer, 1985). Overtone distinguishes six sub-abilities.

A7 Non-literal. Non-literal social communication requires reasoning beyond the literal truth conditions of an utterance and representing what the speaker knows, intends the listener to believe, or expects the listener to recognize. Advanced ToM paradigms therefore distinguish irony and sarcasm, different forms of false statements, humor, and social faux pas (Happé, 1994; Winner & Leekam, 1991). These abilities are particularly relevant to Overtone because vocal realization can substantially alter the intended social and pragmatic interpretation of otherwise identical verbal content.

Table 13: **A1 Emotion.** Fine-grained ToM abilities involving the inference and explanation of others' emotional states.

Sub-ability	Cognitive / ToM Characterization	References
01 Typical emotional reactions	Inferring another person's likely emotion from a situation that typically elicits that reaction, or inferring the likely situation from the person's emotional response. The ability links the meaning of an event to another person's affective state rather than relying solely on surface emotional expression.	(Harris, 1989; Beaudoin et al., 2020)
02 Atypical emotional reactions	Inferring or explaining an emotional response that differs from what would conventionally be expected in the same situation. Successful reasoning requires moving beyond stereotypical event–emotion associations and considering person-specific beliefs, preferences, experiences, or goals.	(Beaudoin et al., 2020)
03 Discrepant emotions	Understanding that different people may experience different emotions toward the same event because they hold different desires, beliefs, evaluations, or social relationships to that event. The same external situation therefore need not imply the same internal affective state across individuals.	(Harris et al., 1989; Beaudoin et al., 2020)
04 Mixed emotions	Understanding that one person may experience more than one emotion simultaneously or successively toward the same event, including emotions of different valence. This requires representing affective states as potentially coexisting rather than mutually exclusive.	(Pons et al., 2004)
05 Hidden emotions	Distinguishing the emotion a person actually experiences from the emotion they intentionally display to others. Correct interpretation requires representing both the internal affective state and the person's motive for masking or controlling its observable expression.	(Harris et al., 1986; Wellman & Liu, 2004)
06 Moral emotions	Understanding emotions arising from evaluation of one's own behavior against moral or social standards, such as guilt following a transgression. The inference links appraisal of one's action and its consequences to the resulting emotional state.	(Beaudoin et al., 2020)
07 Emotion regulation	Understanding that people may intentionally modify, suppress, reappraise, or otherwise regulate their emotional experience or expression. The inference concerns the regulatory goal underlying the difference between an initial emotional response and a later observable state.	(Beaudoin et al., 2020)

C Benchmark Construction and Evaluation

C.1 Expert-Guided Script Generation and Quality Control

Expert-Specified Scenario Constraints. Overtone adopts an expert-guided data construction process in which the social-interaction structure and mental-state configuration of each scenario are explicitly specified before any language-model generation takes place. For each scenario, experts first determine the corresponding ToM task and then construct a scenario blueprint that specifies the event setting, interpersonal relationships, the information available to each character, latent mental states, interaction goals, and the target utterance. These factors are deliberately varied across scenarios in a controlled manner to increase the diversity of social situations represented in Overtone while preserving the core reasoning structure required by the corresponding ToM task. In particular, information access is specified separately for different characters so as to avoid unintended knowledge leakage, while the latent mental states assigned to each character are required to remain consistent with both the observable events and the information available from that character's own perspective. The target utterance must also fit naturally within the surrounding context while avoiding any explicit disclosure of the intended affective or pragmatic interpretation. Through this design, the target

Table 14: **A2 Desire.** Fine-grained ToM abilities involving others' desires, preferences, and desire-dependent behavior.

Sub-ability	Cognitive / ToM Characterization	References
08 Discrepant desires	Understanding that different people can hold different or opposing desires toward the same object, event, or outcome. Successful reasoning requires representing another person's subjective preference even when it differs from one's own.	(Repacholi & Gopnik, 1997; Wellman & Liu, 2004)
09 Desires influence on emotions/actions	Understanding that desires and preferences influence both emotional responses and subsequent behavior. The same outcome may therefore produce different emotions or actions for individuals who want different things.	(Wellman & Woolley, 1990; Harris et al., 1989)
10 Desire-action contradiction	Explaining cases in which a person's observable action appears inconsistent with their expressed desire or preference. Resolving the apparent contradiction may require inferring competing motives, situational constraints, social obligations, or changes in the relative priority of desires.	(Beaudoin et al., 2020)
11 Multiple desires	Understanding that a single person may simultaneously or successively hold multiple desires, including desires that compete with one another. Their eventual behavior may therefore reflect a compromise or prioritization among several desired outcomes.	(Beaudoin et al., 2020)

Table 15: **A3 Intention.** Fine-grained ToM abilities involving the interpretation and prediction of goal-directed action.

Sub-ability	Cognitive / ToM Characterization	References
12 Completion of failed actions	Inferring the intended outcome of another person's unsuccessful or incomplete action. The observer must distinguish what the actor actually achieved from what they were trying to achieve, and may use the inferred intention to complete or predict the unfinished act.	(Meltzoff, 1995)
13 Discrepant intentions	Understanding that similar actions or outcomes may arise from different underlying intentions, or that an observed outcome may differ from what the actor intended. Interpretation therefore requires distinguishing intention from outcome and using contextual information about the actor's goal.	(Killen et al., 2011; Beaudoin et al., 2020)
14 Prediction of actions	Predicting another person's subsequent behavior from their inferred intention or goal. The prediction is based on what the person is trying to accomplish rather than solely on their immediately observable behavior.	(Phillips et al., 2002; Beaudoin et al., 2020)
15 Intentions explanations	Inferring an intention that plausibly explains an observed action or social interaction. This ability is particularly important when the same overt behavior admits several interpretations and its underlying social purpose must be reconstructed from context.	(Beaudoin et al., 2020)

interpretation cannot be determined simply by relying on explicit emotion words or other superficial lexical patterns.

Constrained Candidate Script Generation. Based on the expert-defined scenario blueprint, we use GPT-5.5 to generate candidate scripts under explicit structural and linguistic constraints. Importantly, the language model is used only to realize the predefined scenario structure in natural language and does not determine the core psychological structure of the data. The ToM task type, character-specific information states, intended mental-state configuration, and interaction goals are all specified by experts before generation. During generation, GPT-5.5 is required to preserve these predefined variables while realizing them as natural, concise, and plausible social interactions

Table 16: **A4 Percept.** Fine-grained ToM abilities involving another person's perceptual access and perspective.

Sub-ability	Cognitive / ToM Characterization	References
16 Percept-action link	Understanding that another person's behavior is shaped by what they have perceptually accessed. Predictions must therefore be made from the actor's perceptual perspective rather than from information available only to the observer.	(Beaudoin et al., 2020)
17 Auditory perspective-taking	Representing what another person can or cannot hear and using that representation to infer their likely response, attention, or action. The ability requires distinguishing one's own auditory access from that of another individual.	(Scherer et al., 2015)

Table 17: **A5 Knowledge.** Fine-grained ToM abilities involving epistemic access and another person's state of knowledge.

Sub-ability	Cognitive / ToM Characterization	References
18 Knowledge-pretend play links	Understanding that intentional pretense can depend on knowledge of the object, event, or representation incorporated into the pretend scenario. A person who lacks the relevant knowledge cannot intentionally construct or interpret the corresponding pretense in the same way.	(Beaudoin et al., 2020)
19 Percepts-knowledge links	Understanding that perceptual access, such as seeing or hearing relevant information, can provide knowledge, whereas the absence of such access may leave another person ignorant. Correct inference requires tracking who perceived which information.	(Pillow, 1989; Beaudoin et al., 2020)
20 Information-knowledge links	Understanding that being told, informed, or otherwise given access to information changes what a person can reasonably know. Knowledge should therefore be attributed according to the information available to that person rather than to the observer.	(Wellman & Woolley, 1990; Beaudoin et al., 2020)
21 Knowledge-attention links	Understanding that a person's existing knowledge or familiarity can influence what draws their attention. Information that is novel or unexpected for that person may attract attention differently from information they already know.	(Beaudoin et al., 2020)

grounded in everyday settings. At the same time, the generated content is required to avoid directly explaining the characters' internal mental states, using wording that is overly diagnostic of the intended interpretation, or constructing unnatural dialogue whose primary purpose is to reveal the target answer. For scenarios designed to support multiple vocal realizations, both the textual context and the target utterance are additionally required to remain sufficiently semantically neutral, such that multiple affective or pragmatic interpretations remain plausible when the audio information is unavailable.

Expert Review and Scoring. After candidate generation, all scripts undergo a structured expert review before proceeding to the recording and question-construction stages. This step is designed to verify that each script faithfully instantiates its intended ToM structure while avoiding unintended shortcuts or annotation leakage. As shown in Figure 7, the review protocol combines expert judgment with predefined scoring criteria and acceptance rules, allowing potentially usable candidates to be revised while filtering out scripts that cannot support reliable mental-state inference.

Additional Validation for Shared Scripts. For shared scripts, we perform an additional consistency validation procedure, as these samples are specifically designed to isolate the contribution of vocal realization to mental-state inference. In this setting, different recordings share exactly the same textual context and target utterance, while distinct affective or pragmatic interpretations are induced solely through variations in vocal expression. Experts therefore verify that, in the absence of audio

Table 18: **A6 Belief.** Fine-grained ToM abilities involving false beliefs, higher-order beliefs, and belief-dependent behavior and emotion.

Sub-ability	Cognitive / ToM Characterization	References
22 Content false beliefs	Understanding that a person may hold an incorrect belief about the contents of a familiar object or container when the actual contents differ from what they expect and they have not observed the discrepancy.	(Beaudoin et al., 2020)
23 Location false beliefs	Understanding that a person who does not witness or learn about an object's displacement may continue to believe that it remains at its previous location and may act according to that outdated representation.	(Wimmer & Perner, 1983)
24 Identity false beliefs	Understanding that misleading observable properties can cause another person to form an incorrect belief about an object's or person's identity. Correct reasoning separates the observer's knowledge of the true identity from the target person's representation.	(Beaudoin et al., 2020)
25 Second-order belief	Representing what one person believes about another person's belief, for example, "A thinks that B believes p." Successful reasoning requires recursively tracking the information and perspective available at both levels of the embedded representation.	(Perner & Wimmer, 1985)
26 Beliefs based action/emotions	Understanding or predicting a person's behavior or emotional reaction on the basis of what that person believes, including when the belief is false. The person's subjective representation rather than objective reality determines the expected action or emotion.	(Harris, 1989; Beaudoin et al., 2020)
27 Sequence false beliefs	Understanding a false expectation generated when a familiar or predictable sequence is disrupted by an event that the person does not observe or learn about. The inference requires maintaining the sequence as represented by that person rather than updating it with privileged information.	(Beaudoin et al., 2020)

information, each intended interpretation remains plausible given the same textual context, and that neither the background description nor the target utterance contains explicit cues that uniquely determine one vocal condition. Meanwhile, after incorporating the corresponding vocal realization, the resulting mental-state interpretations must exhibit sufficient discriminability to support meaningful changes in subsequent ToM judgments. Scripts are excluded from the shared subset if their intended interpretation can already be determined from text alone, or if the corresponding vocal realization fails to provide sufficient evidence for the intended interpretation.

Final Consistency Check. Before proceeding to human speech recording and question construction, all retained scripts undergo a final cross-stage consistency review. We re-examine whether the scenario-level ToM task, character-specific information states, latent mental-state configurations, and interaction goals remain mutually consistent after iterative refinement. We further verify that the target utterance continues to satisfy the intended semantic neutrality requirement while remaining suitable for natural human vocal realization. By explicitly separating **expert specification**, **constrained language-model generation**, and **subsequent expert verification**, we minimize the influence of language-model generation biases on the benchmark structure and ensure that the final scenarios are grounded in predefined social-cognitive and mental-state reasoning structures.

C.2 Question and Option Construction

Question Construction Principles. After finalizing each script and its corresponding vocal condition, we manually construct four questions for every sample, following the four progressive evaluation levels in Overtone: Emotional and Attitudinal Understanding (EAU), Mental-State Attribution (MSA), Social-Interaction Reasoning (SIR), and Recursive-Pragmatic Reasoning (RPR). These levels do not represent repeated queries about the same mental state, but instead capture different stages

Table 19: **A7 Non-literal.** Fine-grained ToM abilities involving nonliteral meaning, deception-related intentions, humor, and social appropriateness.

Sub-ability	Cognitive / ToM Characterization	References
28 Irony/sarcasm	Understanding an utterance whose intended evaluative stance differs from its literal content. Interpretation requires recovering the speaker's actual attitude and recognizing that the listener is expected to detect the discrepancy rather than accept the utterance at face value.	(Winner & Leekam, 1991)
29 Egocentric lies	Understanding that a speaker knowingly produces a false statement for a self-serving purpose, such as gaining an advantage or avoiding an undesirable consequence. This requires representing both the speaker's knowledge of the truth and their intention to alter the listener's belief.	(Happé, 1994)
30 White lies	Understanding a deliberately false statement produced primarily to protect another person's feelings or preserve a social relationship. The inference requires distinguishing a prosocial deceptive intention from deception motivated by the speaker's own benefit.	(Happé, 1994; Winner & Leekam, 1991)
31 Involuntary lies	Understanding that a speaker may produce a false statement without intending to deceive because they are mistaken, uninformed, or have forgotten relevant information. The ability distinguishes objective falsehood from intentional deception by representing the speaker's own knowledge state.	(Happé, 1994)
32 Humor	Understanding that a literally false, exaggerated, or otherwise non-serious utterance may be produced for amusement rather than deception. Correct interpretation requires recovering the speaker's playful intention and recognizing that the listener is expected to understand the utterance nonliterally.	(Happé, 1994)
33 Faux pas	Recognizing a socially inappropriate or potentially hurtful utterance produced without the speaker realizing its problematic consequence. Successful reasoning jointly requires representing the speaker's missing knowledge or unintended action and the likely emotional impact on the listener.	(Baron-Cohen et al., 1999)

of social-cognitive inference, ranging from affective interpretation to higher-order reasoning about communicative intentions and recursive mental representations.

EAU evaluates whether a model can infer a speaker's emotional state, attitude, or evaluative stance by integrating social context with vocal expression. MSA further requires reasoning about latent mental states, including knowledge, beliefs, desires, intentions, and concerns. SIR focuses on how these mental states shape ongoing social interactions, including why a speaker adopts a particular expression, what they intend the listener to understand, and what interaction outcome they aim to achieve. RPR further examines whether a listener can reason about a speaker's mental state and communicative intention, including recursive perspective taking, non-literal meanings, and expectations about how an utterance will be interpreted.

Each question is associated with one predefined ToM ability dimension and its corresponding sub-ability. Rather than asking for explicit facts stated in the script, questions require models to infer latent mental states by integrating the social context, character-specific information access, and the realized vocal expression. Therefore, we avoid questions that can be answered through general world knowledge, isolated emotional keywords, local lexical matching, or literal interpretation of the target utterance. For questions involving beliefs, knowledge states, or recursive mental representations, we explicitly specify the relevant character, information source, and temporal state to prevent unintended ambiguity unrelated to ToM reasoning.

Option and Distractor Design. Each question is formulated as a four-way multiple-choice item containing one gold answer and three competing distractors. Unlike randomly generated incorrect

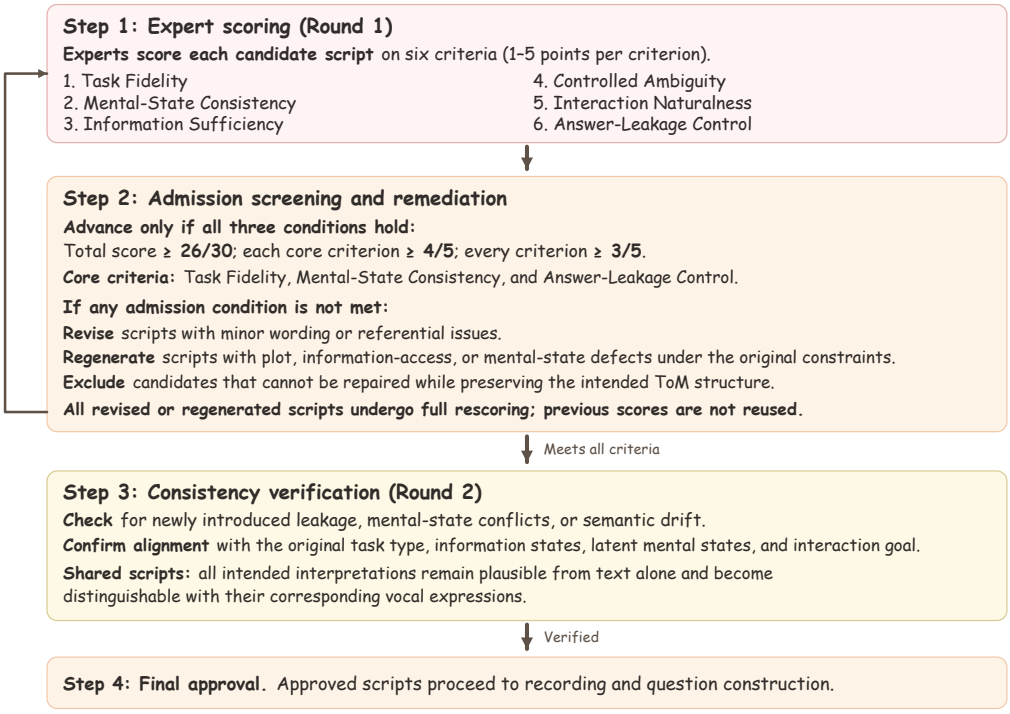

Figure 7: **Expert review and scoring protocol for candidate scripts.** The figure summarizes the expert-based evaluation pipeline applied to all generated script candidates, including the quality dimensions, scoring procedure, acceptance criteria, and the decision paths for revision, regeneration, or exclusion.

options, distractors are designed as plausible alternative mental-state interpretations of the same social interaction. Each distractor preserves part of the scenario information while deviating from the gold answer along a critical psychological dimension, such as incorrect attribution of knowledge states, confusion between desires and beliefs, misinterpretation of communicative intentions, misunderstanding of emotional consequences, or inaccurate representation of another person's mental state.

A valid distractor should satisfy two requirements: (1) it should remain plausible when the target perspective, vocal information, or higher-order mental representation is not fully considered; and (2) it should become distinguishable from the gold answer once all relevant information is integrated. We therefore exclude options that are unrelated to the scenario, logically impossible, introduce nonexistent entities or events, or can be eliminated without ToM reasoning.

To reduce answer-level shortcuts, we further control the surface form of all options. The four alternatives are designed to have comparable length, syntactic complexity, semantic specificity, and explicitness of mental-state descriptions. The gold answer is not systematically longer, more specific, more certain, or richer in psychological terminology than distractors. Answer positions are balanced across the dataset and evaluation levels to avoid positional biases.

Expert Review and Quantitative Scoring. All questions and options undergo structured expert evaluation before being included in the benchmark. Experts independently score each item according to six criteria using a 5-point scale. The criteria and scoring rubric are summarized in Table 20.

Each question is independently evaluated by two experts. The maximum score is 30, and an item is retained only if all of the following conditions are satisfied:

- The total score is at least 26/30;

Table 20: Evaluation criteria and scoring rubric for question and option quality. Each dimension is rated on a 1–5 scale by independent experts.

Criterion	Scoring Guidelines
Ability Alignment	Measures whether the question genuinely evaluates the intended ToM ability. 5: directly targets the intended ability with minimal influence from unrelated skills; 4: clearly aligned with minor auxiliary reasoning; 3: partially involves the target ability but includes noticeable non-ToM factors; 2: the target ability is secondary; 1: does not measure the intended ability.
Gold Support	Measures whether the gold answer is sufficiently supported by the scenario, character information states, and vocal condition. 5: fully supported without additional assumptions; 4: clearly supported with minimal contextual completion; 3: reasonable but relies on implicit assumptions; 2: weakly supported; 1: unsupported by available evidence.
Gold Uniqueness	Measures whether the gold answer can be clearly distinguished from all distractors. 5: uniquely supported and all alternatives are excluded by mental-state reasoning; 4: clearly preferable with minor residual ambiguity; 3: one strong competing interpretation remains; 2: multiple options remain similarly plausible; 1: no unique answer can be established.
Distractor Plausibility	Measures whether distractors represent meaningful competing interpretations. 5: all distractors correspond to distinct psychological errors; 4: all are plausible with minor weakness in distinction; 3: at least one distractor is relatively easy to eliminate; 2: multiple distractors are implausible; 1: distractors can be rejected without understanding the scenario.
Shortcut Control	Measures whether answers can be identified through superficial patterns. 5: no identifiable shortcut; options are well balanced; 4: minor surface differences without predictive value; 3: potential but limited shortcuts exist; 2: obvious answer cues are present; 1: answer can be solved mainly through surface features.
Referential and Temporal Clarity	Measures whether characters, information sources, event order, and mental-state timing are clear. 5: all references and temporal relations are explicit; 4: minor ambiguity without affecting reasoning; 3: limited ambiguity requiring interpretation; 2: ambiguity may alter the answer; 1: the intended mental state cannot be reliably determined.

- The four core dimensions (Ability Alignment, Gold Support, Gold Uniqueness, and Shortcut Control) each receive a score of at least 4;
- No dimension receives a score below 3;
- Both experts agree on the gold answer.

If the two experts disagree by two or more points on any dimension, or disagree on the gold answer, ability assignment, or question validity, a third expert performs adjudication. If a stable interpretation cannot be reached after discussion, the item is removed. Items failing the threshold are revised or discarded rather than being compensated by high scores in other dimensions.

Additional Validation for Shared Scripts. Shared scripts receive additional cross-vocal-condition validation because their design goal is to isolate the contribution of vocal expression to mental-state inference. Since these samples preserve identical textual contexts and target utterances across recordings, differences between conditions should arise from vocal realization rather than textual changes.

Experts evaluate three additional criteria:

- **Textual Compatibility**: without audio, each intended interpretation should remain plausible given the same textual context, and neither the background nor target utterance should uniquely reveal one vocal condition;
- **Vocal Discriminability**: after incorporating the audio, different vocal realizations should support distinct and well-defined mental-state interpretations;

- **Audio–Label Consistency**: the interpretation supported by the actual recording should match the assigned gold answer.

Each criterion is rated on a 1–5 scale, and all three dimensions must achieve a score of at least 4. Scripts are excluded from the shared subset if the intended interpretation can already be determined from text alone or if the corresponding vocal realization fails to provide sufficient support for the intended interpretation.

After human recording, experts repeat this validation procedure using the final audio recordings. Therefore, gold labels for shared scripts are not determined solely by the intended recording condition, but are revalidated against the actual acoustic realization presented to evaluated models.

C.3 Human Speech Recording

All speech data in Overtone are collected from human speakers through a dedicated online recording platform. We adopt human-recorded speech rather than synthesized voices to preserve natural variations in prosody, affective expression, and paralinguistic behaviors, which are often difficult to faithfully reproduce using fixed synthetic speech. The complete recording process consists of four stages: speaker recruitment and screening, pre-recording preparation and instruction, controlled vocal realization, and post-recording quality review.

C.3.1 Speaker Recruitment and Screening

To obtain vocal realizations that are both natural and sufficiently distinguishable across conditions, we adopt a two-stage speaker recruitment and screening procedure. In the initial recruitment stage, we received applications and audition recordings from 100 candidates. The audition materials were evaluated according to predefined criteria, including speech clarity, English proficiency, emotional and prosodic expressiveness, naturalness of delivery, and the ability to follow vocal-expression instructions. Based on the first-stage evaluation, 30 candidates were selected for further assessment.

In the second stage, shortlisted candidates participated in interviews and live auditions. This stage further evaluated their ability to understand the social context of the scripts and the underlying mental states of the characters, as well as their ability to produce distinguishable vocal realizations while keeping the lexical content unchanged. We additionally assessed whether candidates could maintain consistent interpretations of the same character across repeated recordings, together with their availability and reliability throughout the full recording process.

Following the two-stage screening procedure, we selected seven speakers for formal data collection, including three male and four female speakers. All speakers received compensation for their participation.

C.3.2 Recording Preparation and Instructions

Before formal recording, speakers are provided with the corresponding scripts and relevant social-interaction contexts. Rather than asking speakers to mechanically imitate isolated emotion labels, we instruct them to first understand the ongoing interaction, interpersonal relationships, the target speaker's underlying mental state, and communicative stance. Speakers then realize the target utterance in a natural manner that is consistent with the inferred social situation.

For different controlled recordings of the same script, speakers are required to keep the lexical content of the target utterance strictly identical. They are not allowed to add, remove, replace, or paraphrase any words. Under this constraint, speakers can naturally adjust vocal properties such as pitch, loudness, speaking rate, rhythm, and pausing when these variations are necessary to convey the intended vocal condition. We further instruct speakers to maintain a conversational speaking style and avoid excessively exaggerated or theatrical performances, preventing the target vocal condition from becoming unnaturally salient.

Each target utterance is recorded four times. If a recording contains script misreading, unintended lexical modification, substantial background noise, or insufficient realization of the intended vocal condition, the speaker is required to repeat the recording. The final recording is selected by jointly considering audio quality, naturalness of delivery, and the adequacy of the intended vocal realization.

C.3.3 NON-VERBAL VOCALIZATIONS

Beyond conventional prosodic variations, Overtone allows speakers to naturally incorporate non-verbal vocalizations (NVVs) when they are consistent with the social interaction and affective-pragmatic context. These vocal behaviors aim to preserve non-lexical information commonly present in real-world communication but often absent from textual transcriptions. Examples include sighs, light laughter, hesitation sounds, and other non-lexical vocal behaviors that may alter the social meaning conveyed by identical lexical content.

Rather than specifying a fixed inventory or predefined template of NVVs, we provide speakers with limited expressive freedom while requiring them to preserve the core lexical content of the target utterance. The specific type, position, duration, and intensity of NVVs are not strictly predetermined. Instead, speakers naturally realize these behaviors based on their understanding of the character's mental state, interpersonal relationship, and intended affective-pragmatic condition. This design avoids reducing naturally occurring vocal behaviors into a small set of mechanically inserted acoustic markers.

NVVs do not replace or modify the lexical content of the target utterance. Instead, they provide an additional vocal channel that supplements information not explicitly encoded in the transcript. For example, the same utterance may convey substantially different interpersonal attitudes and mental-state interpretations when accompanied by a sigh, restrained laughter, or hesitation. Such samples therefore require models to utilize the complete acoustic signal rather than relying solely on transcription or literal semantic content.

After recording, all samples containing NVVs undergo manual auditory review. Experts examine whether the non-verbal behaviors are consistent with the social context and intended affective-pragmatic condition, naturally integrated with the surrounding speech, and do not compromise the intelligibility of the lexical content. NVVs that are overly exaggerated, contextually inappropriate, or disruptive to speech understanding result in re-recording or exclusion of the corresponding version.

C.3.4 ONLINE RECORDING PLATFORM

All formal recordings are collected through a dedicated online recording platform developed for Overtone. The platform presents the scenario background, target utterance, and recording instructions within a unified interface, ensuring a consistent recording procedure across speakers and recording sessions. The contextual information and target utterance are presented separately, allowing speakers to first understand the social interaction before producing the corresponding vocal realization.

For controlled recordings of the same script, the platform keeps the lexical content of the target utterance unchanged while providing different vocal realization requirements according to the assigned condition. The platform also supports multiple recording attempts and subsequent quality management, allowing invalid recordings to be identified and removed before entering the final benchmark.

Figure 8 illustrates the interface of the online recording platform and the workflow used during speech collection.

During recording, if a sample contains script-reading errors, unintended lexical changes, substantial background noise, or fails to realize the target vocal condition, it is marked for re-recording rather than directly included in the final dataset.

C.3.5 POST-RECORDING QUALITY REVIEW

All recorded utterances undergo a structured quality-control procedure before being included in Overtone. We first perform two mandatory validity checks: **Lexical Fidelity**, which requires the spoken content to match the target utterance without unintended additions, omissions, substitutions, or paraphrasing; and **Technical Integrity**, which excludes recordings with severe background noise, clipping, truncation, overlapping speech, or other artifacts that may interfere with speech perception. Naturally occurring NVVs are permitted as long as they do not alter or obscure the lexical content. Recordings that fail either check are rejected and re-recorded when necessary.

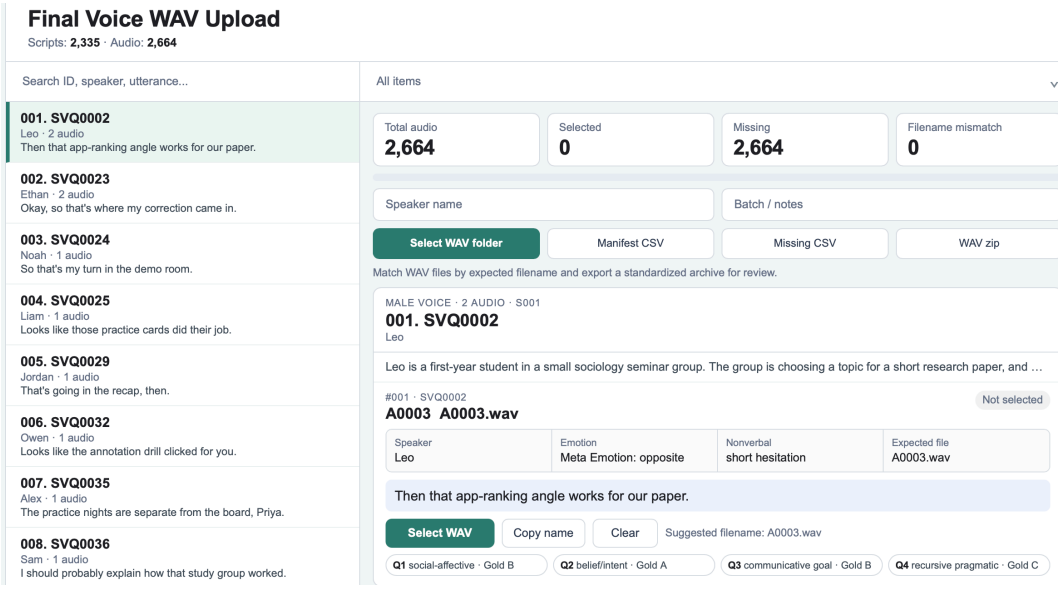

Figure 8: **Online recording platform for human speech collection.** The platform provides speakers with scenario context, target utterances, and recording instructions, while supporting controlled vocal realization and multiple recording attempts.

Recordings that pass the basic checks are further evaluated for **Audio Clarity**, **Naturalness**, **Vocal-Condition Fidelity**, and **Contextual Appropriateness**; samples containing NVVs are additionally assessed for **Paralinguistic Integration**. As illustrated in Figure 9, the review combines independent human ratings, predefined acceptance thresholds, and adjudication for inconsistent judgments. When multiple takes satisfy all requirements, the final recording is selected by jointly considering naturalness, fidelity to the intended vocal condition, and contextual appropriateness, rather than simply choosing the most acoustically salient realization.

Additional Validation for Shared Recordings. For shared scripts, recordings that pass individual quality control are further reviewed jointly across vocal conditions. Experts verify that the lexical content remains identical, that each realization is individually natural and valid, and that the intended difference in affective or pragmatic interpretation is supported by vocal information rather than by recording artifacts or unintended textual variation. After the final recordings are selected, experts revalidate the correspondence between the actual audio and the associated annotations. Recordings that do not reliably support the intended interpretation are re-recorded or excluded rather than retaining the originally assigned label by construction.

C.4 Evaluation Metrics

We evaluate model performance at three complementary levels: *question-level correctness*, *cross-question mental-state consistency*, and *vocal and multimodal sensitivity*. Let $\mathcal{D} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N}$ denote the evaluation set, where each sample x_i is associated with four questions and $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,4})$ denotes the corresponding gold answers. For modality condition $m \in \{T, A, A{+}T\}$, we denote the model prediction for question q of sample i by $\hat{y}_{i,q}^{m}$, where T, A, and $A{+}T$ refer to Text-only, Audio-only, and Audio+Text, respectively.

Accuracy (Acc). We use question-level accuracy as the basic evaluation metric. For modality condition m, accuracy is defined as

$$\mathrm{Acc}_m = \frac{1}{4N} \sum_{i=1}^{N} \sum_{q=1}^{4} \mathbb{I}\left[\hat{y}_{i,q}^{m} = y_{i,q}\right], \tag{1}$$

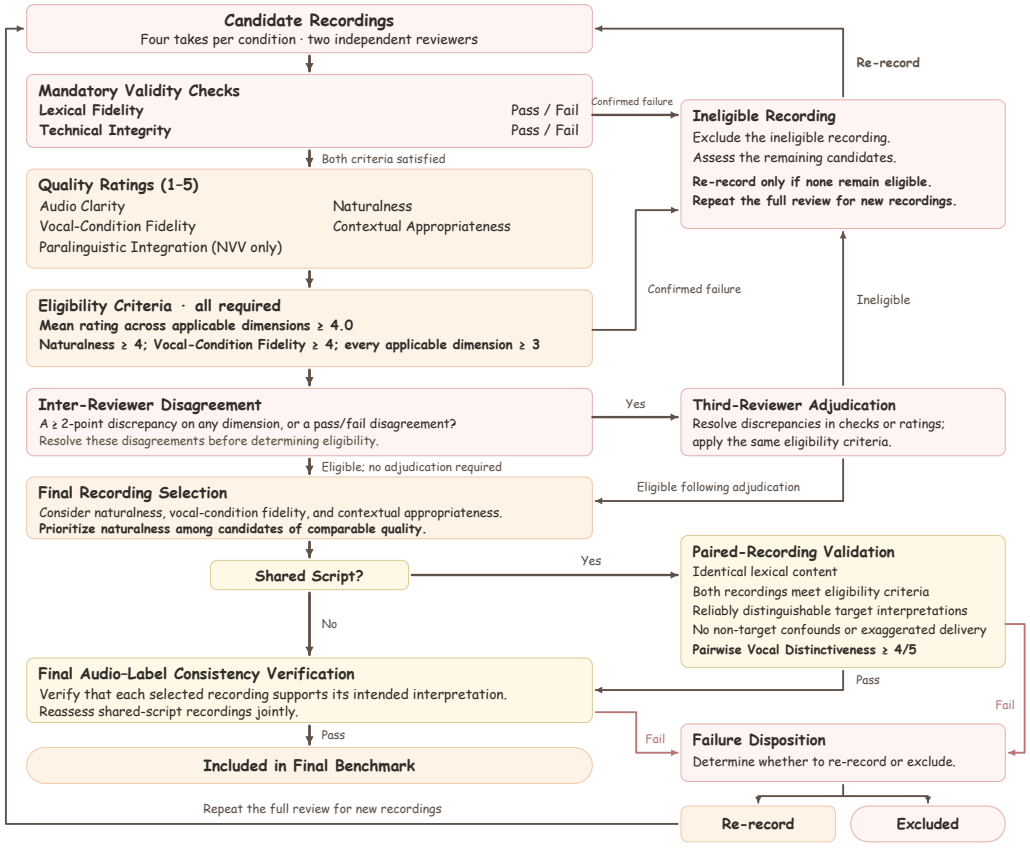

Figure 9: **Post-recording quality-control protocol.** The pipeline summarizes mandatory validity checks, quantitative quality assessment, acceptance thresholds, reviewer adjudication, and the re-recording or exclusion of samples that fail to satisfy the predefined criteria. Shared scripts additionally undergo pairwise validation to verify that distinct vocal realizations support the intended differences in mental-state interpretation while preserving identical lexical content.

where $\mathbb{I}[\cdot]$ is the indicator function. We additionally report accuracy separately for the four question levels, EAU, MSA, SIR, and RPR, as well as for the affective–pragmatic and ToM ability dimensions.

Exact Match over Four Questions (EM4). Question-level accuracy does not indicate whether a model maintains a coherent interpretation of the same social interaction across multiple related mental-state questions. We therefore report EM4, which counts a sample as correct only when all four associated questions are answered correctly:

$$\mathrm{EM4} = \frac{1}{N}\sum_{i=1}^{N}\prod_{q=1}^{4}\mathbb{I}\left[\hat{y}_{i,q} = y_{i,q}\right]. \tag{2}$$

EM4 therefore provides a stricter measure of cross-question consistency than average accuracy: a locally correct answer is insufficient unless the model preserves a compatible mental-state interpretation across all four reasoning levels.

Pair Accuracy (PairAcc). A distinctive component of Overtone consists of *shared scripts*, in which the textual context and target utterance are held fixed while the vocal realization is systematically varied. This controlled setting allows us to test whether a model can distinguish mental states that are expressed through voice rather than lexical content.

Let $\mathcal{P}_\Delta$ denote the set of predefined matched question pairs whose two instances share the same textual input and question but differ in vocal realization and gold interpretation. For a pair $p = (i, j, q) \in \mathcal{P}_\Delta$, PairAcc is defined as

$$\mathrm{PairAcc} = \frac{1}{|\mathcal{P}_\Delta|} \sum_{(i,j,q)\in\mathcal{P}_\Delta} \mathbb{I}\left[\hat{y}_{i,q} = y_{i,q} \ \wedge \ \hat{y}_{j,q} = y_{j,q}\right]. \tag{3}$$

Unlike ordinary accuracy, PairAcc requires the model to correctly resolve *both* members of a controlled vocal contrast. It therefore penalizes models that obtain reasonable overall accuracy while collapsing different vocal realizations onto the same mental-state interpretation.

Flip Precision (FP). PairAcc measures whether both members of a vocal contrast are correct, but does not directly characterize the reliability of a model's *prediction changes*. We therefore further measure Flip Precision. A prediction flip occurs when the model changes its answer across two matched vocal realizations: $\hat{y}_{i,q} \neq \hat{y}_{j,q}$. FP evaluates how often such a change corresponds to the correct change in mental-state interpretation:

$$\mathrm{FP} = \frac{\sum_{(i,j,q)\in\mathcal{P}_\Delta} \mathbb{I}\left[\hat{y}_{i,q} \neq \hat{y}_{j,q}\right] \mathbb{I}\left[\hat{y}_{i,q} = y_{i,q} \wedge \hat{y}_{j,q} = y_{j,q}\right]}{\sum_{(i,j,q)\in\mathcal{P}_\Delta} \mathbb{I}\left[\hat{y}_{i,q} \neq \hat{y}_{j,q}\right]}. \tag{4}$$

Thus, FP is a conditional measure of the *correctness of model revisions*. A model may frequently change its predictions across vocal conditions while still obtaining low FP if these changes do not track the corresponding gold mental-state distinction. Conversely, high FP indicates that prediction changes induced by vocal variation are systematically aligned with the intended interpretation.

PairAcc and FP are not reported for Text-only evaluation. Under this condition, the two members of a shared pair provide identical textual input and contain no information identifying the vocal realization; consequently, pairwise voice-sensitive metrics are not meaningful for assessing Text-only models.

Vocal-Cue Gain. To quantify the contribution of vocal information beyond the transcript, we define Vocal-Cue Gain (VCG) as the change in accuracy from Text-only to Audio+Text:

$$\mathrm{VCG} = \mathrm{Acc}_{A+T} - \mathrm{Acc}_T. \tag{5}$$

VCG is reported in percentage points (pp). Positive values indicate that access to the speech signal improves prediction beyond the textual input, whereas negative values indicate that incorporating vocal information degrades performance. The same definition is applied within individual affective, pragmatic, and ToM ability subsets when analyzing dimension-specific vocal effects.

Best Unimodal Performance. Although VCG measures the contribution of speech relative to Text-only input, it does not distinguish genuine multimodal integration from cases where one unimodal channel is already substantially stronger than the other. We therefore define the best unimodal performance as

$$B = \max\left(\mathrm{Acc}_T, \mathrm{Acc}_A\right). \tag{6}$$

Unimodal Oracle Complementarity. We next estimate the amount of complementary correctness exposed by the two unimodal predictions. Flattening all question instances into $\{1, \ldots, M\}$, where $M = 4N$, we construct an oracle that counts an instance as solvable whenever *either* Text-only or Audio-only produces the correct answer:

$$O = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[\left(\hat{y}_i^T = y_i\right) \vee \left(\hat{y}_i^A = y_i\right)\right]. \tag{7}$$

The interval $O - B$ therefore measures the complementary performance margin available from the union of correct unimodal predictions. Importantly, O is an oracle with respect to the *observed unimodal prediction sets*, rather than a theoretical upper bound on what a jointly conditioned multimodal model can achieve.

Fusion Gain (FG). We define Fusion Gain as the absolute improvement of Audio+Text over the stronger unimodal condition:

$$\mathrm{FG} = \mathrm{Acc}_{A+T} - B. \tag{8}$$

FG isolates the net benefit of joint multimodal input from improvements that could already be obtained by simply selecting the stronger single modality. Accordingly, $\mathrm{FG} > 0$ indicates a genuine gain over both unimodal settings, whereas $\mathrm{FG} < 0$ indicates multimodal interference.

Complementarity Utilization (CU). FG measures absolute improvement but does not account for how much complementary information is actually available for a particular model. We therefore define Complementarity Utilization as

$$\mathrm{CU} = \frac{\mathrm{Acc}_{A+T} - B}{O - B}, \qquad O > B. \tag{9}$$

CU normalizes the multimodal gain by the model-specific complementarity margin $O - B$. It therefore measures the fraction of complementary correctness exposed by the two unimodal predictions that is converted into actual Audio+Text performance. When $B \leq \mathrm{Acc}_{A+T} \leq O$, CU lies in $[0, 1]$: $\mathrm{CU} = 0$ indicates that joint input provides no gain over the best unimodal condition, whereas $\mathrm{CU} = 1$ indicates that the model recovers the full unimodal-oracle complementarity margin. Negative values indicate multimodal interference. In principle, $\mathrm{CU} > 1$ is also possible when joint multimodal reasoning correctly resolves instances missed by both unimodal runs.

Together, these metrics separate three aspects that are conflated by aggregate accuracy: whether a model is correct on individual questions (Acc), whether it maintains a consistent interpretation across related mental-state queries (EM4), and whether it both responds appropriately to vocal variation (PairAcc and FP) and effectively exploits complementary information across modalities (VCG, FG, and CU). Unless otherwise specified, Acc, EM4, PairAcc, FP, and the oracle performance O are reported as percentages, VCG and FG as percentage-point differences, and CU as a dimensionless ratio.

D Additional Experimental Details

D.1 Human Experiment

D.1.1 Participants and Recruitment

We recruit 48 participants through a dedicated online evaluation platform to establish human performance on Overtone. Participants are required to have sufficient ability to understand English instructions, social scenarios, target utterances, and multiple-choice questions. Before the formal evaluation, participants receive detailed instructions explaining the task format and response procedure.

All participants complete the evaluation remotely through the dedicated platform. Each participant is compensated according to the number of assigned questions. All collected responses are recorded anonymously and used only for research purposes.

D.1.2 Evaluation Interface and Procedure

We develop a dedicated online evaluation interface for human assessment. As shown in Figure 10, the interface presents each question as an independent sample and provides participants with the corresponding social context, target utterance, and audio input when applicable, followed by a multiple-choice question with candidate answers.

Participants are instructed to select the option that best reflects the underlying mental state of the speaker or character according to the provided information. No additional explanations or intermediate reasoning are required.

To prevent information leakage across related items, each question is presented independently. Participants do not evaluate multiple question levels associated with the same audio sample. For shared scripts containing multiple vocal realizations, different audio versions derived from the same script

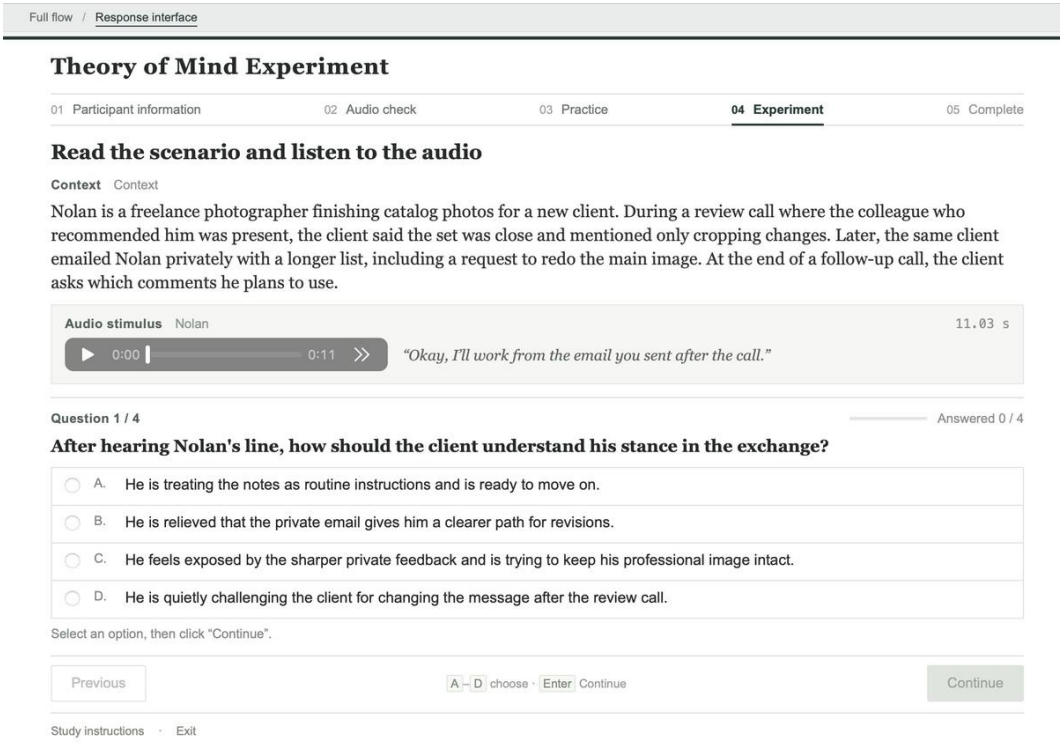

Figure 10: Screenshot of the dedicated online evaluation interface for human assessment. The interface presents the social context, target utterance, audio input, and multiple-choice questions as independent evaluation samples.

are never assigned to the same participant. This assignment strategy prevents participants from exploiting cross-question consistency or directly comparing different vocal expressions, ensuring that each judgment reflects independent interpretation based only on the provided sample.

D.1.3 Human Judgment Assignment and Quality Control

Each evaluation question is independently answered by three different participants. We adopt a redundant independent judgment strategy to reduce individual variability and obtain reliable human performance estimates.

The evaluation procedure consists of three stages: (1) sample assignment, where questions are randomly assigned to participants while satisfying the independence constraints described above; (2) human judgment, where participants complete assigned questions through the online evaluation interface; and (3) response aggregation, where independent judgments are collected and compared with the predefined gold labels.

For each question, the final human response is obtained by aggregating the three independent judgments. Human performance is then calculated using the same evaluation metrics applied to model predictions, including accuracy and exact-match accuracy where applicable.

D.2 Experimental Setup

D.2.1 Evaluated Models

We evaluate 16 representative multimodal models on Overtone. These models cover speech-specialized language models, general-purpose omni-modal models, and human-centric multimodal systems. Detailed descriptions of each evaluated model are provided below.

GPT-Audio-1.5. GPT-Audio-1.5 OpenAI (2026) is a speech-language model designed for native audio interaction. It directly processes spoken inputs and generates language outputs, enabling end-to-end speech understanding and reasoning.

Audio Flamingo-Next. Audio Flamingo-Next Ghosh et al. (2026) extends the Audio Flamingo series toward stronger audio perception and reasoning capabilities through large-scale audio-language alignment.

MOSS-Audio-8B. MOSS-Audio-8B Yang et al. (2026) is an open-source speech-language model focusing on speech understanding and conversational interaction.

Step-Audio-R1.5. Step-Audio-R1.5 Zhang et al. (2026) is a reasoning-oriented speech model designed to improve complex spoken-language understanding through enhanced reasoning capabilities.

Kimi-Audio. Kimi-Audio KimiTeam et al. (2025) is an audio-language model supporting general speech understanding and multimodal interaction.

MiMo-Audio. MiMo-Audio Xiaomi LLM-Core Team (2025) is an audio-centric multimodal model designed for speech perception, instruction following, and reasoning over spoken inputs.

Audio-Reasoner. Audio-Reasoner Xie et al. (2025) is a reasoning-focused audio-language model developed to improve structured reasoning over auditory information.

Gemini-3.7-Flash. Gemini-3.7-Flash Google DeepMind (2026b) is a general-purpose multimodal model supporting native processing of text, vision, and audio inputs.

Gemini-3.1-Pro. Gemini-3.1-Pro Google DeepMind (2026a) is a high-capability multimodal foundation model designed for complex reasoning across multiple modalities, including audio.

Qwen3.5-Omni-Plus. Qwen3.5-Omni-Plus Qwen Team (2026) is an omni-modal large language model integrating text, vision, and audio understanding within a unified framework.

MiMo-V2.5. MiMo-V2.5 Xiaomi MiMo (2026) is a general multimodal model designed for advanced perception and reasoning across multiple input modalities.

MiniCPM-o-4.5. MiniCPM-o-4.5 Cui et al. (2026) is an open-source omni-modal model supporting audio, vision, and text interaction.

Gemma. Gemma Gemma Team (2026) is a multimodal language model family designed for efficient multimodal understanding and generation.

Nemotron-3-Nano-Omni. Nemotron-3-Nano-Omni NVIDIA (2026) is an omni-modal model designed for efficient multimodal reasoning and interactive applications.

HumanOmni V2. HumanOmni V2 Yang et al. (2025) is a human-centric multimodal model designed to improve perception and reasoning in human-centered scenarios.

OmniSapiens 2.0. OmniSapiens 2.0 Ong et al. (2026) is a unified multimodal model targeting human-centered intelligence through integrated perception, reasoning, and interaction.

All models are evaluated in a zero-shot setting without additional training or task-specific adaptation.

E Fine-Grained Results on Overtone

We report Audio+Text accuracy across the eight task dimensions, 19 affective–pragmatic dimensions, and 33 ToM subabilities. The model order and evaluation setting follow the main results.

E.1 Task Dimensions

Table 21: Fine-grained results across eight task dimensions. Audio+Text accuracy (%).

Model	Task Dimension							
	UOT	**SIT**	**PST**	**FBT**	**AST**	**HT**	**SST**	**FPR**
Random Guess	24.88	25.70	23.45	25.00	24.79	27.53	25.35	30.00
Human	94.33	90.65	91.80	90.13	91.91	92.74	93.00	88.75
Speech Large Language Models (SpeechLLMs)								
GPT-Audio-1.5	71.67	66.47	69.74	68.85	67.70	75.00	71.19	65.00
Audio-Flamingo-Next	68.35	66.24	67.57	62.94	64.99	68.89	66.13	62.50
MOSS-Audio-8B	68.35	66.12	64.71	66.67	63.10	70.39	61.70	66.25
Step-Audio-R1.5	67.12	64.02	65.40	64.08	62.15	67.28	65.07	64.38
Kimi-Audio	59.85	59.00	61.07	60.44	58.36	63.59	63.48	60.62
MiMo-Audio	63.92	60.98	62.23	60.44	56.67	61.98	57.98	53.75
Audio-Reasoner	57.88	53.15	57.97	54.94	54.58	55.30	58.87	53.75
Omni Large Language Models (OmniLLMs)								
Gemini-3.7-Flash	81.40	74.53	80.80	77.83	77.86	81.80	85.11	78.12
Gemini-3.1-Pro	79.56	73.95	80.73	77.99	76.91	81.22	85.11	79.38
Qwen3.5-Omni-Plus	75.62	70.44	70.90	70.95	70.68	76.61	74.20	75.00
MiMo-V2.5	70.94	65.89	66.49	68.12	65.29	74.19	70.83	71.25
MiniCPM-o-4.5	67.36	68.22	65.87	66.59	64.38	69.47	62.41	65.00
Gemma-4-E4B	64.41	64.84	65.25	61.49	63.45	67.05	71.19	69.38
Nemotron-3-Nano-Omni	67.86	66.12	64.01	64.81	63.13	66.59	59.49	63.12
Human-Centric Omni Large Language Models (HC-Omni LLMs)								
HumanOmni V2	60.96	60.05	56.27	56.07	53.37	53.92	47.16	50.62
OmniSapiens 2.0	58.87	58.88	54.26	54.29	52.14	52.88	39.27	48.12
PEAR (Ours)	**87.32**	**83.29**	**89.47**	**87.06**	**87.20**	**88.82**	**91.49**	**85.00**

E.2 Affective and Pragmatic Dimensions

Table 22: Fine-grained results across 19 affective-pragmatic dimensions.

Model	Basic						Social			
	Hap	**Sad**	**Ang**	**Bad**	**Sur**	**Fea**	**Gui**	**Sha**	**Emb**	**Pri**
Random Guess	24.56	26.82	24.12	25.00	27.46	24.05	25.37	25.89	25.12	23.51
Human	93.36	93.18	90.49	93.92	94.20	92.14	92.04	89.80	94.43	90.42
Speech Large Language Models (SpeechLLMs)										
GPT-Audio-1.5	73.23	62.27	48.01	71.85	80.58	74.52	70.77	65.43	73.39	80.35
Audio-Flamingo-Next	59.29	56.82	43.36	74.77	71.21	70.48	63.43	63.52	62.87	83.21
MOSS-Audio-8B	75.66	51.36	32.08	61.71	78.12	63.81	63.93	54.85	62.38	85.20
Step-Audio-R1.5	64.60	52.05	39.38	65.77	69.64	65.00	66.17	55.74	63.61	79.98
Kimi-Audio	62.17	61.82	54.20	66.67	64.73	66.90	57.21	64.67	67.95	62.19
MiMo-Audio	66.37	49.09	31.19	54.73	69.20	58.81	57.96	47.19	57.80	79.60
Audio-Reasoner	49.12	51.59	37.83	57.21	64.06	58.10	56.34	55.23	63.49	65.42
Omni Large Language Models (OmniLLMs)										
Gemini-3.7-Flash	78.32	75.68	67.04	89.86	87.28	80.95	77.11	76.53	77.97	87.44
Gemini-3.1-Pro	72.57	74.09	67.70	87.61	89.73	83.57	76.00	78.70	80.82	83.21
Qwen3.5-Omni-Plus	79.20	60.68	46.46	74.77	88.39	71.90	71.02	60.46	71.04	88.18
MiMo-V2.5	67.04	56.59	43.14	68.24	80.36	70.24	68.28	58.55	71.91	82.34
MiniCPM-o-4.5	74.34	54.32	34.07	62.84	79.69	64.76	62.81	58.16	64.60	88.43
Gemma-4-E4B	55.75	61.82	51.11	71.40	65.40	74.76	65.67	65.94	69.80	78.23
Nemotron-3-Nano-Omni	68.81	53.86	32.30	61.71	76.79	62.86	59.58	57.53	61.76	84.58
Human-Centric Omni Large Language Models (HC-Omni LLMs)										
HumanOmni V2	75.66	42.05	22.57	43.69	64.51	49.29	50.75	42.09	53.09	82.59
OmniSapiens 2.0	76.77	36.36	21.68	39.64	62.28	48.57	49.63	38.01	47.52	84.33
PEAR (Ours)	**90.27**	**86.59**	**74.78**	**94.82**	**95.09**	**90.95**	**84.33**	**90.43**	**87.62**	**94.78**

Table 22 (continued): Fine-grained results across 19 affective-pragmatic dimensions.

Model	Mixed	Higher-order Pragmatic Attitudes							
	Ind	Adm	Iro	Sin	Opp	EQ	EA	MN	Neg
Random Guess	26.85	19.23	25.23	25.00	26.41	28.08	23.61	23.93	26.09
Human	93.15	90.73	92.09	92.66	89.44	94.86	88.19	88.93	89.49
Speech Large Language Models (SpeechLLMs)									
GPT-Audio-1.5	75.28	74.83	67.66	59.62	59.86	60.62	73.26	50.36	71.74
Audio-Flamingo-Next	77.96	68.53	64.11	50.52	67.78	57.88	63.89	58.21	67.03
MOSS-Audio-8B	77.13	73.60	58.49	59.27	60.92	61.99	68.40	60.00	64.13
Step-Audio-R1.5	75.46	67.83	64.56	56.29	66.37	53.08	60.07	53.57	65.58
Kimi-Audio	61.11	48.25	60.89	46.85	74.65	38.36	47.22	55.00	58.70
MiMo-Audio	68.98	75.70	49.54	68.01	47.71	39.38	74.31	46.79	64.86
Audio-Reasoner	59.72	53.50	55.50	45.45	64.79	36.99	49.31	60.00	43.84
Omni Large Language Models (OmniLLMs)									
Gemini-3.7-Flash	81.67	75.70	94.61	59.97	91.90	76.03	56.25	60.36	79.71
Gemini-3.1-Pro	82.78	75.87	92.78	56.99	89.61	74.32	56.60	58.57	79.71
Qwen3.5-Omni-Plus	78.89	81.64	73.85	71.15	59.15	70.89	75.69	47.86	78.26
MiMo-V2.5	73.70	71.33	70.41	66.78	56.16	69.86	71.53	38.93	77.54
MiniCPM-o-4.5	75.93	74.83	56.88	64.16	54.58	55.82	78.82	55.71	68.48
Gemma-4-E4B	68.52	53.67	74.77	46.68	80.81	41.78	40.97	61.07	56.16
Nemotron-3-Nano-Omni	75.19	71.50	55.16	70.28	54.05	67.12	70.49	39.64	74.64
Human-Centric Omni Large Language Models (HC-Omni LLMs)									
HumanOmni V2	63.80	69.41	34.63	67.66	39.08	44.86	76.74	55.71	55.07
OmniSapiens 2.0	60.83	73.43	25.34	**75.35**	28.87	42.81	**85.07**	50.71	53.99
PEAR (Ours)	**93.15**	**86.71**	**96.79**	67.83	**98.24**	**81.51**	67.36	**65.36**	**81.88**

E.3 THEORY-OF-MIND SUBABILITIES

Table 23: Fine-grained results across 33 ToM ability dimensions.

Model	Percepts		Knowledge				Beliefs					
	P1	P2	K1	K2	K3	K4	B1	B2	B3	B4	B5	B6
Random Guess	27.78	27.43	57.14	36.36	23.83	29.41	19.05	11.11	0.00	24.30	26.54	22.22
Human	90.91	92.72	71.43	81.82	91.97	97.06	95.24	88.89	100.00	92.99	92.89	88.89
Speech Large Language Models (SpeechLLMs)												
GPT-Audio-1.5	69.70	74.27	**100.00**	72.73	71.62	79.41	80.95	44.44	77.78	78.62	72.73	77.78
Audio-Flamingo-Next	62.12	68.69	71.43	54.55	69.75	67.65	57.14	66.67	77.78	71.26	64.79	55.56
MOSS-Audio-8B	63.89	67.23	57.14	**90.91**	65.86	67.65	66.67	66.67	66.67	67.76	67.61	77.78
Step-Audio-R1.5	57.58	66.75	**100.00**	72.73	65.19	76.47	66.67	66.67	66.67	66.47	63.32	**88.89**
Kimi-Audio	51.77	60.68	71.43	36.36	56.89	61.76	61.90	55.56	55.56	62.50	61.86	77.78
MiMo-Audio	55.81	60.44	85.71	72.73	65.46	70.59	71.43	66.67	66.67	63.67	62.07	77.78
Audio-Reasoner	59.34	47.57	42.86	54.55	57.43	67.65	52.38	**77.78**	66.67	53.86	59.67	66.67
Omni Large Language Models (OmniLLMs)												
Gemini-3.7-Flash	79.29	82.28	85.71	72.73	80.19	85.29	90.48	**77.78**	**100.00**	79.91	78.58	77.78
Gemini-3.1-Pro	78.28	83.50	**100.00**	**90.91**	80.05	**88.24**	95.24	55.56	88.89	78.74	78.16	**88.89**
Qwen3.5-Omni-Plus	67.42	72.57	**100.00**	72.73	74.83	**88.24**	80.95	**77.78**	88.89	77.22	73.67	77.78
MiMo-V2.5	67.68	68.69	**100.00**	72.73	68.81	67.65	80.95	55.56	88.89	73.60	70.85	77.78
MiniCPM-o-4.5	66.67	67.48	85.71	81.82	68.54	70.59	76.19	**77.78**	88.89	73.95	67.19	**88.89**
Gemma-4-E4B	64.14	69.42	**100.00**	72.73	67.34	50.00	76.19	55.56	55.56	70.44	64.79	77.78
Nemotron-3-Nano-Omni	62.12	64.56	85.71	63.64	66.40	67.65	90.48	66.67	66.67	69.63	66.98	**88.89**
Human-Centric Omni Large Language Models (HC-Omni LLMs)												
HumanOmni V2	56.82	55.83	85.71	72.73	63.19	70.59	47.62	55.56	88.89	56.07	56.95	77.78
OmniSapiens 2.0	55.81	51.94	71.43	54.55	61.98	73.53	38.10	44.44	77.78	57.48	54.55	55.56
PEAR (Ours)	**88.89**	**88.11**	85.71	**90.91**	**89.16**	85.29	**100.00**	66.67	**100.00**	**88.79**	**88.19**	**88.89**

Table 23 (continued): Fine-grained results across 33 ToM ability dimensions. Panel (b) of 3. Audio+Text accuracy (%).

Model	Desires				Emotions						
	D1	D2	D3	D4	E1	E2	E3	E4	E5	E6	E7
Random Guess	26.76	33.08	31.75	31.58	31.25	9.09	16.67	25.13	23.86	18.10	24.82
Human	91.76	93.23	95.62	85.96	100.00	72.73	83.33	90.91	92.40	91.38	88.65
Speech Large Language Models (SpeechLLMs)											
GPT-Audio-1.5	75.00	75.19	67.15	75.44	56.25	45.45	83.33	43.85	56.12	72.41	64.54
Audio-Flamingo-Next	66.76	78.95	60.58	71.93	68.75	72.73	83.33	62.03	58.09	64.66	78.72
MOSS-Audio-8B	71.18	79.70	63.87	77.19	68.75	63.64	91.67	61.50	49.33	56.90	71.63
Step-Audio-R1.5	70.88	75.94	64.96	66.67	75.00	72.73	66.67	62.03	48.17	56.90	73.76
Kimi-Audio	67.65	69.17	57.66	66.67	**100.00**	45.45	50.00	32.09	58.45	39.66	51.77
MiMo-Audio	65.00	68.42	60.58	64.91	31.25	36.36	58.33	36.90	43.25	44.83	68.09
Audio-Reasoner	58.53	64.66	51.46	52.63	93.75	45.45	83.33	56.68	57.19	60.34	68.79
Omni Large Language Models (OmniLLMs)											
Gemini-3.7-Flash	82.06	75.19	74.45	89.47	75.00	63.64	91.67	60.96	73.82	79.31	82.98
Gemini-3.1-Pro	81.76	77.44	71.53	85.96	93.75	54.55	83.33	59.36	72.65	79.31	76.60
Qwen3.5-Omni-Plus	79.41	79.70	69.34	78.95	87.50	63.64	91.67	50.27	59.16	68.10	80.85
MiMo-V2.5	74.12	76.69	67.15	73.68	62.50	54.55	75.00	49.73	53.89	68.10	72.34
MiniCPM-o-4.5	71.18	75.19	65.69	75.44	81.25	63.64	66.67	48.13	49.60	52.59	74.47
Gemma-4-E4B	71.76	69.17	57.30	66.67	50.00	54.55	83.33	37.43	53.80	56.90	68.09
Nemotron-3-Nano-Omni	67.94	76.69	65.33	77.19	75.00	54.55	75.00	48.13	48.61	55.17	72.34
Human-Centric Omni Large Language Models (HC-Omni LLMs)											
HumanOmni V2	60.29	70.68	55.11	49.12	43.75	63.64	83.33	32.62	42.54	36.21	71.63
OmniSapiens 2.0	58.24	70.68	53.65	56.14	43.75	54.55	58.33	30.48	37.62	31.03	66.67
PEAR (Ours)	**93.24**	**87.22**	**83.58**	**92.98**	93.75	**90.91**	**100.00**	**85.03**	**83.29**	**90.52**	**95.74**

Table 23 (continued): Fine-grained results across 33 ToM ability dimensions.

Model	Intentions				Non-Literal Communication					
	I1	I2	I3	I4	N1	N2	N3	N4	N5	N6
Random Guess	18.18	26.67	22.30	24.63	24.51	33.33	19.72	33.33	26.67	31.03
Human	81.82	88.89	91.99	92.06	90.68	91.67	91.55	100.00	91.67	96.55
Speech Large Language Models (SpeechLLMs)										
GPT-Audio-1.5	81.82	74.22	77.87	72.86	62.04	66.67	70.42	58.33	61.67	58.62
Audio-Flamingo-Next	63.64	71.11	70.21	66.34	63.36	66.67	59.15	75.00	63.33	68.97
MOSS-Audio-8B	81.82	74.22	70.73	69.69	59.41	50.00	67.61	66.67	56.67	51.72
Step-Audio-R1.5	63.64	71.56	69.34	70.31	60.19	75.00	64.79	58.33	61.67	55.17
Kimi-Audio	81.82	66.22	60.80	63.92	58.82	58.33	67.61	58.33	58.33	65.52
MiMo-Audio	72.73	69.78	68.64	65.67	50.51	66.67	66.20	58.33	63.33	48.28
Audio-Reasoner	63.64	60.00	52.26	54.42	53.32	41.67	45.07	66.67	66.67	31.03
Omni Large Language Models (OmniLLMs)										
Gemini-3.7-Flash	90.91	84.44	82.93	79.95	80.51	83.33	80.28	91.67	81.67	82.76
Gemini-3.1-Pro	90.91	87.11	82.58	80.09	79.20	75.00	74.65	**100.00**	81.67	79.31
Qwen3.5-Omni-Plus	81.82	78.67	78.92	76.08	67.84	66.67	69.01	75.00	61.67	72.41
MiMo-V2.5	81.82	72.00	74.74	70.92	63.12	75.00	67.61	91.67	53.33	75.86
MiniCPM-o-4.5	72.73	72.00	75.61	69.83	58.22	75.00	67.61	83.33	65.00	48.28
Gemma-4-E4B	54.55	72.89	72.30	67.42	63.36	58.33	63.38	83.33	63.33	58.62
Nemotron-3-Nano-Omni	72.73	72.89	71.60	68.32	57.38	83.33	73.24	50.00	58.33	62.07
Human-Centric Omni Large Language Models (HC-Omni LLMs)										
HumanOmni V2	54.55	63.11	67.77	59.43	42.56	83.33	52.11	75.00	55.00	34.48
OmniSapiens 2.0	63.64	56.89	66.03	59.91	37.54	66.67	60.56	33.33	55.00	31.03
PEAR (Ours)	**100.00**	**92.89**	**90.94**	**88.42**	**84.34**	**91.67**	**84.51**	**100.00**	**91.67**	**96.55**

F Audio-Only and Text-Only Results

F.1 Audio-Only Results

Table 24: Direct results under the Audio-only input condition (%).

Model	All Samples						Shared Pairs
	EAU↑	MSA↑	SIR↑	RPR↑	Avg↑	EM4↑	PairAcc↑
Speech Large Language Models (SpeechLLMs)							
GPT-Audio-1.5	60.62	72.41	72.11	75.56	70.18	36.79	4.32
Audio-Flamingo-Next	63.18	64.15	66.74	70.20	66.07	25.38	7.51
MOSS-Audio-8B	60.55	68.36	68.17	68.62	66.42	26.84	7.85
Step-Audio-R1.5	54.05	62.61	65.77	65.65	62.02	21.88	7.11
Kimi-Audio	56.34	56.76	58.18	58.26	57.39	14.64	**13.76**
MiMo-Audio	52.25	61.41	58.37	63.85	58.97	19.56	4.92
Audio-Reasoner	58.22	58.18	50.64	51.54	54.65	9.53	1.80
Omni Large Language Models (OmniLLMs)							
Gemini-3.7-Flash	**76.13**	**79.05**	80.56	80.44	**79.04**	54.65	10.37
Gemini-3.1-Pro	74.21	78.79	**80.71**	**81.01**	78.68	**54.88**	10.17
Qwen3.5-Omni-Plus	70.65	76.35	78.57	78.12	75.92	47.22	7.85
MiMo-V2.5	60.40	69.11	68.21	70.68	67.10	31.87	7.91
MiniCPM-o-4.5	56.91	67.30	69.97	72.75	66.73	28.34	2.26
Gemma-4-E4B	54.95	65.95	66.59	68.99	64.12	27.89	1.00
Nemotron-3-Nano-Omni	54.32	66.25	66.29	66.67	63.38	23.91	6.12
Human-Centric Omni Large Language Models (HC-Omni LLMs)							
HumanOmni V2	47.26	58.45	54.54	59.61	54.96	15.09	5.32
OmniSapiens 2.0	43.51	57.28	54.50	54.77	52.52	15.02	5.52

F.2 TEXT-ONLY RESULTS

Table 25: Direct results under the Text-only input condition (%).

Model	All Samples						Shared Pairs
	EAU↑	MSA↑	SIR↑	RPR↑	Avg↑	EM4↑	PairAcc↑
Speech Large Language Models (SpeechLLMs)							
GPT-Audio-1.5	57.28	70.27	70.83	74.06	68.11	33.56	0.80
Audio-Flamingo-Next	60.36	64.11	66.89	69.52	65.22	25.23	0.00
MOSS-Audio-8B	53.00	67.34	65.99	66.74	63.27	24.40	0.00
Step-Audio-R1.5	52.33	63.66	67.30	66.22	62.38	23.95	5.72
Kimi-Audio	45.12	61.07	59.16	58.82	56.04	19.41	0.00
MiMo-Audio	50.04	62.09	60.81	65.28	59.55	20.16	**5.98**
Audio-Reasoner	57.55	59.12	53.08	50.19	54.98	8.90	0.00
Omni Large Language Models (OmniLLMs)							
Gemini-3.7-Flash	64.83	75.23	75.68	74.85	72.64	45.31	1.13
Gemini-3.1-Pro	**66.03**	**76.39**	**76.46**	**77.10**	**74.00**	**47.64**	0.86
Qwen3.5-Omni-Plus	58.71	71.36	73.16	73.65	69.22	39.34	1.00
MiMo-V2.5	55.82	68.39	70.35	70.42	66.24	32.62	5.72
MiniCPM-o-4.5	52.29	66.25	68.17	69.82	64.13	27.85	0.00
Gemma-4-E4B	52.89	64.79	66.93	68.66	63.32	26.43	0.00
Nemotron-3-Nano-Omni	53.34	64.04	66.29	64.83	62.12	26.46	4.99
Human-Centric Omni Large Language Models (HC-Omni LLMs)							
HumanOmni V2	44.07	57.73	55.22	56.79	53.45	16.10	0.00
OmniSapiens 2.0	40.80	57.47	56.46	55.44	52.54	17.49	0.00

F.3 AUDIO+TEXT VERSUS TEXT-ONLY: FINE-GRAINED COMPARISONS

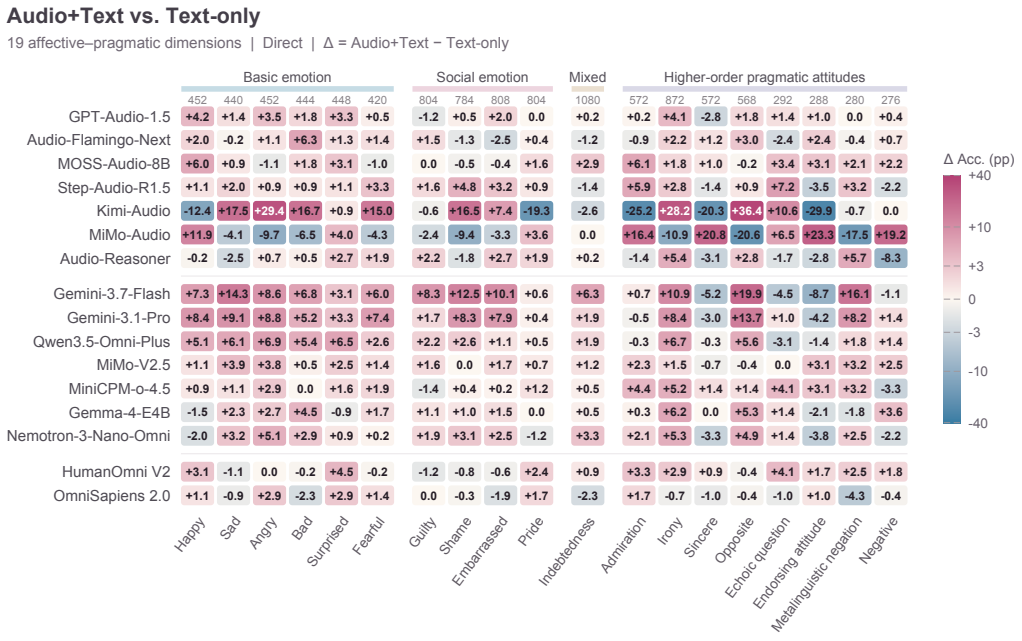

Figure 11: Accuracy differences between Audio+Text and Text-only across affective and pragmatic categories. Values are Audio+Text minus Text-only accuracy in percentage points; positive values indicate gains from adding audio.

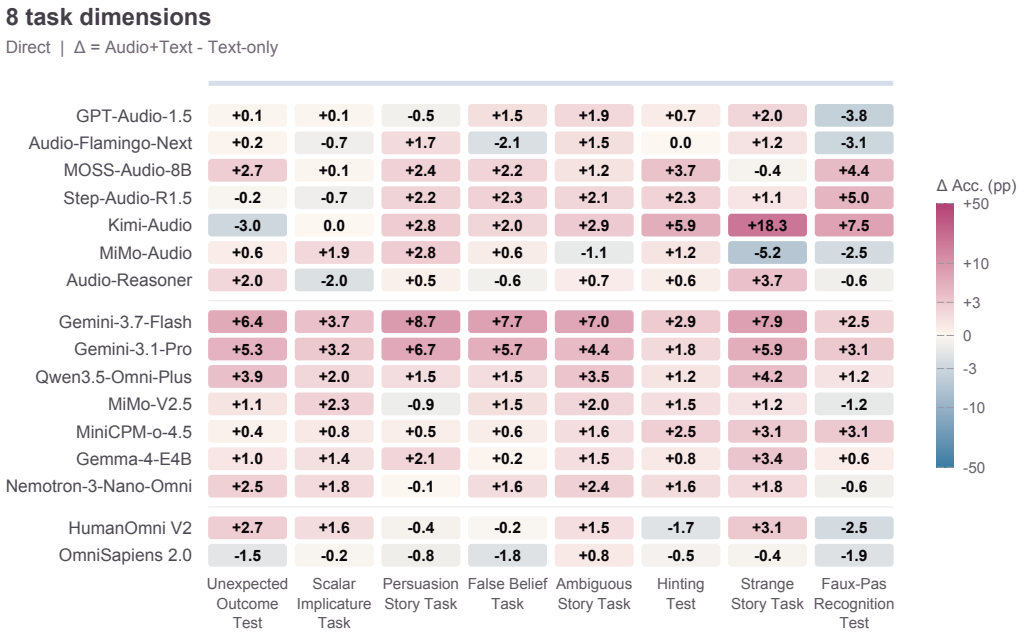

Figure 12: Audio+Text minus Text-only accuracy across eight task categories, in percentage points.

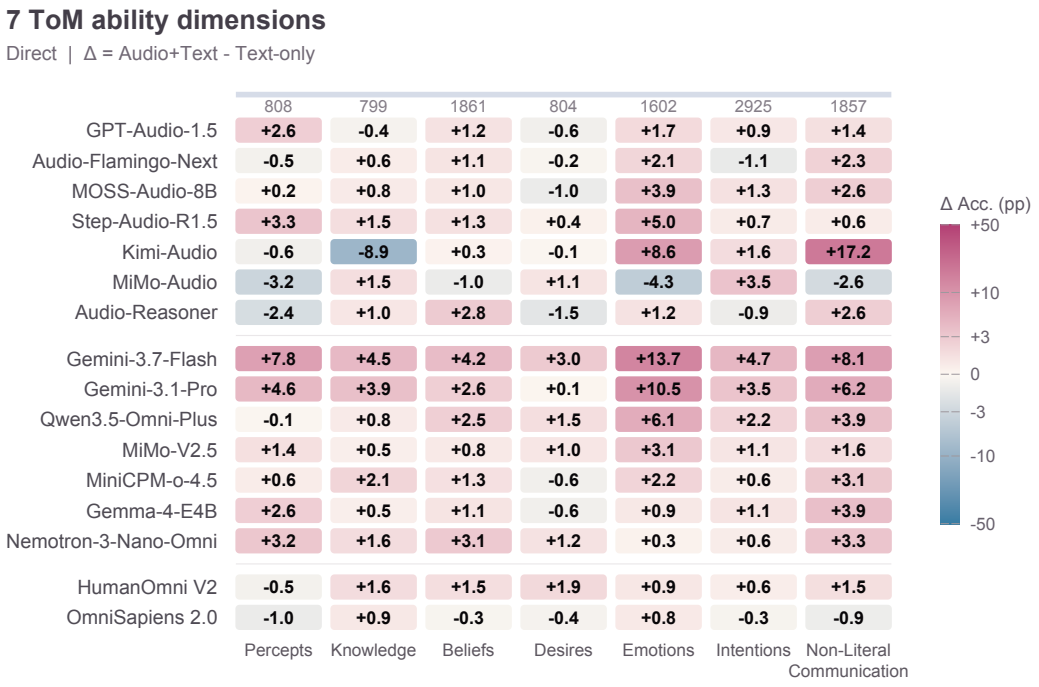

Figure 13: Audio+Text minus Text-only accuracy across seven ToM ability dimensions, in percentage points.

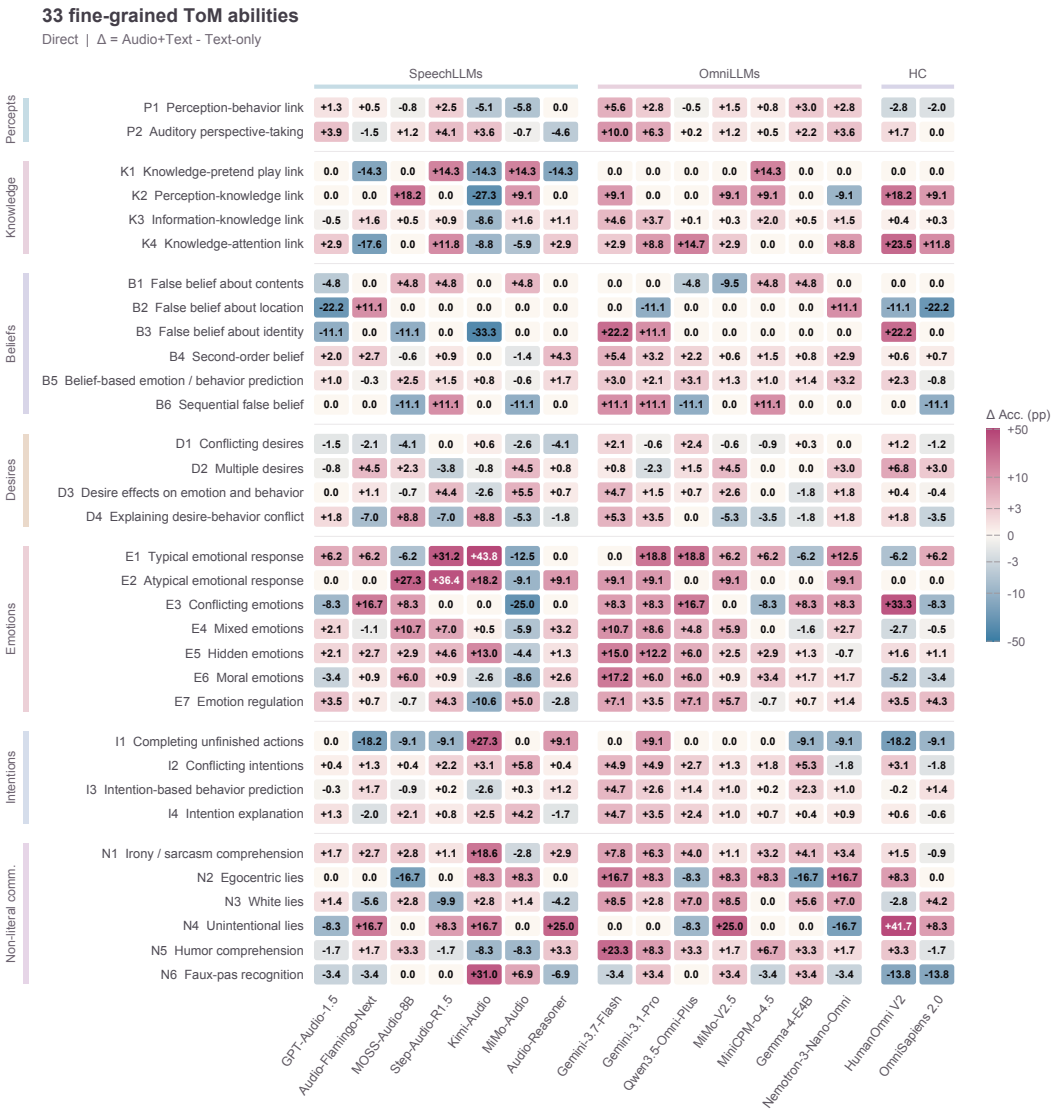

Figure 14: Audio+Text minus Text-only accuracy across 33 fine-grained ToM abilities, in percentage points.

G Complete Results of ToM CoT and PC-ToM

Table 26: Complete ToM CoT results on Overtone across three input conditions. All values are percentages.

	All Samples						Shared Pairs	
Backbone	**EAU**	**MSA**	**SIR**	**RPR**	**Avg**	**EM4**	**PairAcc**	**Flip Precision**
Text-only								
GPT-Audio-1.5	55.44	68.81	67.57	69.97	65.45	28.94	1.40	25.30
Qwen3.5-Omni-Plus	57.21	71.85	73.80	72.86	68.93	39.26	1.13	40.48
Gemini-3.1-Pro	61.04	73.99	75.19	74.59	71.20	42.57	0.93	20.29
Gemini-3.7-Flash	62.46	75.60	75.71	74.44	72.05	44.74	1.40	29.17
Audio-only								
GPT-Audio-1.5	61.34	71.02	71.32	72.60	69.07	34.53	8.51	46.21
Qwen3.5-Omni-Plus	74.10	78.90	80.18	81.01	78.55	51.99	11.17	56.76
Gemini-3.1-Pro	76.01	79.92	81.53	81.79	79.81	57.43	11.84	56.33
Gemini-3.7-Flash	77.52	80.59	81.01	81.42	80.13	57.55	11.84	56.33
Audio+Text								
GPT-Audio-1.5	60.06	70.12	69.37	70.91	67.61	31.68	6.98	43.39
Qwen3.5-Omni-Plus	68.77	74.74	77.78	78.72	75.00	47.82	4.92	56.06
Gemini-3.1-Pro	76.01	79.62	81.46	81.42	79.63	57.55	12.50	60.06
Gemini-3.7-Flash	76.58	80.03	81.83	81.46	79.97	57.43	11.17	54.02

Table 27: Complete PC-ToM results on Overtone with Audio+Text input. All values are percentages.

	All Samples						Shared Pairs	
Backbone	**EAU**	**MSA**	**SIR**	**RPR**	**Avg**	**EM4**	**PairAcc**	**Flip Precision**
Audio+Text								
GPT-Audio-1.5	56.49	63.18	63.81	65.24	62.18	21.02	3.72	21.96
Qwen3.5-Omni-Plus	62.24	68.06	69.52	70.05	67.46	29.80	4.45	39.41
Gemini-3.1-Pro	74.32	75.23	76.20	76.80	75.64	44.97	10.11	54.87
Gemini-3.7-Flash	75.56	75.45	76.50	77.82	76.33	44.33	9.44	47.18

H PEAR ADAPTATION ACROSS DIVERSE THEORY-OF-MIND BENCHMARKS

To evaluate the generalization of PEAR beyond Overtone, we instantiate the framework on three external Theory-of-Mind benchmarks: MuMA-ToM, MMToM-QA, and Hi-ToM. Rather than transferring an identical prompting strategy across datasets, PEAR follows a unified predictive-coding inference principle with dataset-specific observation adapters. The underlying reasoning process remains unchanged: constructing latent mental-state hypotheses, predicting observable consequences, evaluating prediction errors, and selectively updating interpretations when observations contradict prior predictions.

H.1 UNIFIED PREDICTIVE-CODING FORMULATION

Given an input instance

$$X = (C, Q, O),$$

where C denotes contextual information, Q denotes the query with candidate answers, and O denotes observable evidence, PEAR first constructs an initial mental-state hypothesis:

$$H_0 = f_\theta(C, Q),$$

where f_θ denotes a frozen language model backbone.

The current hypothesis generates expected observable consequences:

$$\hat{O} = g(H_0, C).$$

The discrepancy between predicted and observed consequences is formulated as:

$$e = \Delta(O, \hat{O}),$$

where the evidence relation is categorized into three states:

$$e \in \{\text{support}, \text{contradict}, \text{unresolved}\}.$$

Specifically, *support* indicates that observations provide positive evidence consistent with the current hypothesis; *contradict* indicates that observations contain evidence incompatible with the predicted outcome; and *unresolved* indicates that the available observations are insufficient to confirm or reject the hypothesis.

PEAR performs selective hypothesis revision:

$$H_1 = \begin{cases} \text{Update}(H_0, e), & e = \text{contradict}, \\ H_0, & e \in \{\text{support}, \text{unresolved}\}. \end{cases}$$

The final prediction is obtained through a deterministic mapping:

$$y = \text{Mapper}(H_1, Q).$$

This formulation ensures that hypothesis revision is driven only by explicit prediction violations, preventing unsupported absence from being interpreted as negative evidence.

H.2 MuMA-ToM: Hybrid Predictive Reasoning Adapter

For MuMA-ToM, PEAR operates on temporally grounded textual event representations. Given a scenario description S and associated action events A, we first construct a fused representation:

$$S_f = \mathrm{Fusion}(S, A).$$

The initial mental-state hypothesis is then generated as:

$$H_0 = f_\theta(S_f, Q).$$

Since MuMA-ToM contains heterogeneous ToM reasoning categories, PEAR adopts different transition adapters for different question families:

$$\mathcal{A}_{\mathrm{MuMA}} = \{\mathcal{A}_B, \mathcal{A}_{BG}, \mathcal{A}_{SG}\},$$

corresponding to belief reasoning, belief-of-goal reasoning, and social-goal reasoning, respectively.

For belief reasoning, PEAR maintains a recurrent mental-state trace:

$$R_{t+1} = \mathrm{Revise}(R_t, o_{t+1}, \Delta(o_{t+1}, \hat{o}_{t+1})),$$

where R_t denotes the current latent mental-state representation.

For goal-oriented reasoning, PEAR models consistency between latent goals and observed state transitions:

$$\tau : (s_t, g_t, a_t) \to (s_{t+1}, g_{t+1}),$$

where s_t denotes the world state, g_t denotes the latent goal state, and a_t denotes the observed action.

The final hybrid inference follows:

$$H^* = \begin{cases} H_{\mathrm{V5}}, & \mathrm{Valid}(H_{\mathrm{V5}}) = 1, \\ H_{\mathrm{QC}}, & \mathrm{otherwise}. \end{cases}$$

H.3 MMToM-QA: Prediction-Error Overlay

For MMToM-QA, PEAR is instantiated as a prediction-error refinement module over an atomic state-transition reasoning process.

For each candidate answer o_i, PEAR constructs an option-conditioned hypothesis:

$$H_i = (M_i, A_i),$$

where M_i denotes the implied mental state and A_i denotes the associated behavioral trajectory.

The hypothesis predicts future state transitions:

$$\hat{\tau}_i = T(H_i).$$

Given the observed trajectory:

$$\tau = \{z_1, z_2, \ldots, z_n\},$$

the prediction discrepancy is measured as:

$$\epsilon_i = D(\tau, \hat{\tau}_i).$$

The prediction-error overlay modifies the original inference only when a clear contradiction is identified:

$$o^* = \begin{cases} \arg\min_i \epsilon_i, & \epsilon_i > \delta, \\ o_{\text{base}}, & \epsilon_i \leq \delta. \end{cases}$$

H.4 HI-TOM: PERSPECTIVE-AWARE ADAPTER

Hi-ToM primarily evaluates perspective-dependent knowledge reasoning, where information accessibility plays a central role. Therefore, PEAR is evaluated with and without an explicit perspective adapter.

Without structured perspective control, PEAR directly performs predictive reasoning:

$$H_1 = \text{PEAR}(C, Q, O).$$

To incorporate perspective-dependent accessibility constraints, we introduce a visibility adapter:

$$O' = \text{Filter}(O, V),$$

where V represents the visibility and knowledge-access constraints of the target agent.

The perspective-aware inference is formulated as:

$$H_1' = \text{PEAR}(C, Q, O').$$

The perspective adapter only modifies the observation space available for prediction, while the underlying predictive-error-driven hypothesis revision mechanism remains unchanged.

I ADDITIONAL EXPERIMENTAL RESULTS

I.1 HI-TOM FULL RESULTS

Table 28: Fine-grained accuracy on Hi-ToM (%).

Method	**Reasoning order**					**Overall**
	O0	**O1**	**O2**	**O3**	**O4**	**Avg**
SimToM	**100.00**	**77.50**	60.00	60.00	57.50	71.00
BIP-ALM	10.00	17.50	10.00	20.00	15.00	14.50
LIMP	5.00	10.00	7.50	2.50	7.50	6.50
AutoToM	95.00	75.00	70.00	67.50	55.00	72.50
PEAR	**100.00**	75.00	35.00	45.00	40.00	59.00
PEAR + Perspective Taking	**100.00**	72.50	**92.50**	**100.00**	**100.00**	**93.00**

I.2 MUMA-TOM FULL RESULTS

Table 29: Fine-grained accuracy on MuMA-ToM (%).

Method	Mental-state dimension			Overall
	Belief	Social goal	Belief of goal	Avg
SimToM	54.60	43.50	44.80	47.63
BIP-ALM	41.20	34.10	30.60	33.90
LIMP	93.40	67.70	68.70	76.60
AutoToM	88.33	77.00	**79.00**	81.44
PEAR	**98.67**	**99.00**	75.67	**91.11**

I.3 MMToM-QA Full Results

Table 30: Fine-grained accuracy on MMToM-QA (%).

Method	Mental-state dimension		Overall
	Belief	Goal	Avg
SimToM	75.67	26.33	51.00
BIP-ALM	64.33	48.00	56.17
LIMP	60.00	50.67	55.33
AutoToM	**96.67**	69.33	83.00
PEAR	83.00	**84.00**	**83.50**

PEAR by fine-grained question type (data-file codes)

Method	Belief			Goal			
	1.1	1.2	1.3	2.1	2.2	2.3	2.4
PEAR	97.00	82.00	70.00	98.67	82.67	80.00	74.67

J DATASET EXAMPLES

BASIC EMOTION

happy

Task: Unexpected Outcome Test

BACKGROUND

Maya's birthday dinner at her apartment has just finished. Her friend Ben had chosen the framed group photo that everyone signed, but he was stuck covering a late shift, so Nora carried it in and gave the toast. After the last guest leaves, Ben arrives and gives Maya his birthday wish in person. One slice of cake is still on the counter.

TARGET UTTERANCE — Maya

So you made it while there was still cake.



QUESTIONS & ANSWERS

EAU — Answer: A

Emotion / Typical emotional reactions

After hearing Maya's line, how should Ben understand her attitude toward his late arrival?

- **A.** She is genuinely glad he came and is treating the lateness lightly.
- **B.** She is hurt that he missed most of the dinner but is keeping the reply controlled.
- **C.** She is unsure whether he came because he felt obligated rather than because he wanted to.
- **D.** She is politely acknowledging the visit without much personal investment.

MSA — Answer: C

Desire / Desires influence on emotions/actions

What is Maya more likely hoping will happen next between them?

- **A.** She hopes Ben will notice that the timing still stings and offer a more personal repair.
- **B.** She hopes to end the birthday topic quickly and keep the rest of the night routine.
- **C.** She hopes they can share a small friendly moment and move past the late timing.
- **D.** She hopes the credit for the signed photo can be sorted out without making things awkward.

SIR — Answer: B

Intention / Intentions explanations

What is Maya most likely trying to do by saying this to Ben?

- **A.** Keep the exchange brief while leaving Ben room to respond.
- **B.** Welcome his arrival in a playful, low-pressure way.
- **C.** Signal that the timing mattered without directly accusing him.
- **D.** Move the attention back to who handled the group gift.

RPR — Answer: D

Intention / Prediction of actions

After hearing the line, what should Ben revise about what Maya wants to do next?

- **A.** He should think she wants a quieter apology before acting as if everything is settled.
- **B.** He should think she wants him to explain the group photo before they talk casually.
- **C.** He should think she wants to avoid any one-on-one conversation tonight.
- **D.** He should think she is open to an easy, friendly moment despite the late arrival.

BASIC EMOTION

surprised

Task: Ambiguous Story Task

BACKGROUND

Evan's car is being repaired, so Chris, a coworker who passes his park-and-ride, has been driving him to the office this week. Evan has a Thursday interview for a coordinator spot on Chris's project, but he has not brought it up during the rides. As Chris parks, he mentions that the manager asked him to sit on that interview panel.

TARGET UTTERANCE — Evan

So you're on that panel now.



QUESTIONS & ANSWERS

EAU — Answer: C

Emotion / Typical emotional reactions

After hearing this line, how should Chris understand Evan's state?

A. He is treating the update as ordinary and keeping the ride conversation neutral.
B. He is irritated that Chris did not mention his panel role before now.
C. He is caught off guard by the link between the rides and the interview.
D. He is pleased that someone he knows will be involved in the interview.

MSA — Answer: B

Knowledge / Information-knowledge links

Which aspect of the situation is Evan most likely taking in as he reacts?

A. He is treating Chris's comment as an invitation to discuss the interview with him.
B. He is noticing that the ride favor now overlaps with his role in the interview.
C. He is reading the announcement as a hint that Chris expects repayment for the rides.
D. He is focusing on whether Thursday's ride schedule will have to change.

SIR — Answer: D

Intention / Intentions explanations

What is Evan most likely trying to do by responding this way?

A. He wants to thank Chris while keeping the interview out of the conversation.
B. He wants to end the ride arrangement before it becomes awkward.
C. He wants to challenge Chris to explain why he accepted the panel role.
D. He wants to mark the new connection and give Chris a chance to respond.

RPR — Answer: B

Knowledge / Information-knowledge links

After hearing Evan's line, what should Chris most reasonably pick up on?

A. Chris should assume he is simply confirming a scheduling detail.
B. Chris should notice that his announcement has changed how the favor may look to him.
C. Chris should think he is asking him to help him in the interview.
D. Chris should take the line as a request to keep the rides secret.

SOCIAL EMOTION

guilty

Task: Persuasion Story Task

BACKGROUND

Lena is a first-year clarinetist in the school band. Marcus, the senior section leader, agreed to stay after rehearsal to help her with a difficult two-bar entrance. Marcus had said he needed to catch the 5:20 bus, but the practice took longer because Lena kept asking to try the entrance closer to full speed. Marcus has just started closing his case, and Lena has noticed the clock.

TARGET UTTERANCE — Lena

We can leave that entrance for tomorrow.



QUESTIONS & ANSWERS

EAU — **Answer: C**

Emotion / Moral emotions

After hearing this line, how should Marcus understand Lena's state?

A. She is pleased that the entrance finally sounds acceptable and wants the progress noticed.
B. She thinks Marcus's approach has slowed her down and is holding back irritation.
C. She realizes her request may have cost Marcus time and is trying to soften that.
D. She is worried that stopping now will leave her unprepared and is masking it.

MSA — **Answer: A**

Desire / Discrepant desires

What does Lena most likely want to happen next?

A. She wants Marcus to feel free to leave without doing more coaching tonight.
B. She wants Marcus to stay long enough to get the entrance up to tempo.
C. She wants Marcus to admit that her faster practice idea was reasonable.
D. She wants Marcus to reassure her that the unfinished entrance is not serious.

SIR — **Answer: D**

Intention / Intentions explanations

What is Lena most likely trying to do by saying this?

A. To make the stopping point sound routine and keep the exchange smooth.
B. To prompt Marcus to offer one clearer practice method before he leaves.
C. To close the session as if the run-through proved enough for now.
D. To release Marcus from staying longer without spelling out an apology.

RPR — **Answer: B**

Intention / Prediction of actions

After hearing this, how should Marcus revise his read of Lena's likely next move?

A. She is likely to ask for one more try if he sounds willing.
B. She is likely to pack up and avoid taking more of his time.
C. She is likely to question the slow practice method before leaving.
D. She is likely to mark the run as a success and leave upbeat.

SOCIAL EMOTION

shame

Task: False Belief Task

BACKGROUND

Roommates Maya and Ellie share one narrow bathroom sink. Maya usually wipes the basin after brushing her teeth while Ellie is still asleep, but she often leaves her face wash and mug on the counter. Ellie has just finished clearing the counter and placing Maya's things in a small bin beside the sink. When Maya walks in, Ellie points to the bin and says the area finally looks usable.

TARGET UTTERANCE — Maya

I didn't realize it looked like that from your side.



QUESTIONS & ANSWERS

EAU — Answer: C

Emotion / Moral emotions

After hearing Maya's line, how should Ellie understand Maya's state?

A. She has accepted the new setup and is trying to keep things easy.
B. She is hurt that her regular wiping seems to have gone unnoticed.
C. She feels exposed because the counter makes her look careless.
D. She is irritated that the sink area was changed without asking.

MSA — Answer: A

Knowledge / Percepts-knowledge links

Which part of the moment does Maya seem to be using as her main cue?

A. Her own toiletries have been presented as the clutter Ellie noticed.
B. The cleared sink will make both roommates' morning routine easier.
C. Ellie's organizing effort may be meant as a fair contribution to the room.
D. Ellie may be waiting to see whether Maya will challenge the new setup.

SIR — Answer: D

Desire / Multiple desires

What is Maya most likely trying to do with this response?

A. She wants to approve the arrangement so they can leave the topic alone.
B. She wants Ellie to recognize she has also been maintaining the sink.
C. She wants to object to the move without starting a direct argument.
D. She wants to keep some dignity while admitting how she came across.

RPR — Answer: B

Intention / Prediction of actions

After hearing the line, what should Ellie think Maya is most likely trying to do next?

A. Accept the bin system for now and avoid another discussion about the sink.
B. Move her things on her own so she is not seen that way again.
C. Ask to split the counter differently before treating the change as settled.
D. Describe her wiping routine so Ellie understands her part of the upkeep.

SOCIAL EMOTION

pride

Task: Ambiguous Story Task

BACKGROUND

During a statistics recitation, TA Jordan could not get the class past a flawed solution on the board. Nora, a student in the section, had written a short counterexample on scratch paper but stayed quiet while Jordan called on Sam, a senior student who often helps. Nora slid the paper to Sam; Sam restated it to the room, and Jordan used it to finish the explanation. After class, Jordan notices Nora's handwriting and says Sam told him the counterexample began with her note.

TARGET UTTERANCE — Nora

So that scratch-paper example made it to the board.



QUESTIONS & ANSWERS

EAU — Answer: B

Emotion / Typical emotional reactions

After hearing this line, how should Jordan understand Nora's reaction?

A. She is bothered that Sam's version drew attention before hers did.
B. She is taking quiet satisfaction in seeing her idea recognized.
C. She is uneasy about being connected to the example at all.
D. She is giving a routine answer with little personal investment.

MSA — Answer: D

Desire / Multiple desires

What is Nora most likely hoping will happen next?

A. She wants Jordan to move on before more attention lands on her.
B. She wants Sam to continue getting credit for the explanation.
C. She wants Jordan to challenge Sam about how he used the note.
D. She wants Jordan to notice her contribution without making a scene.

SIR — Answer: A

Intention / Intentions explanations

What social purpose does Nora most likely have in saying this line?

A. To mark her part in the solution while staying modest about it.
B. To distance herself from the explanation in case it was wrong.
C. To signal that Jordan should stop mentioning the sheet.
D. To push Jordan toward confronting Sam about credit.

RPR — Answer: C

Desire / Desires influence on emotions/actions

After hearing this line, what should Jordan expect Nora is likely to do next in their class relationship?

A. Offer fewer ideas so she avoids being put on the spot.
B. Ask for a public announcement about the example next time.
C. Stay open to helping if Jordan gives her role some quiet credit.
D. Let Sam speak for her because the arrangement works better.

MIXED SOCIAL EMOTION

Task: Scalar Implicature Task

indebtedness

BACKGROUND

The campus art club has just closed a small print exhibition. Maya, the club captain, spent the afternoon moving folding display panels from storage while Noah, the backup presenter, greeted visitors after the scheduled host lost her voice. Several guests thanked Noah for keeping the opening running smoothly. As the room empties, Maya stacks the last panel and Noah pauses beside the exit.

TARGET UTTERANCE

Noah

You put a lot into those panels today.



QUESTIONS & ANSWERS

EAU

Answer: B

Emotion / Mixed emotions

After hearing this line, how should Maya understand Noah's attitude toward what happened?

A. He is mostly pleased that the exhibition gave him a visible role.

B. He appreciates her help but feels the weight of owing her.

C. He is irritated that she handled the setup without asking him.

D. He is making a neutral closing comment so they can both leave.

MSA

Answer: A

Knowledge / Information-knowledge links

After hearing the line, what information does Noah seem to be holding in mind?

A. He recognizes that Maya's hidden setup work helped him receive visible credit.

B. He assumes Maya already knows the visitors valued her work just as much.

C. He thinks the display setup was separate from how the opening was received.

D. He is still unsure whether Maya did more work than the rest of the club.

SIR

Answer: C

Intention / Intentions explanations

What is Noah most likely trying to accomplish by saying this now?

A. He wants to end the cleanup with a simple practical comment.

B. He wants to explain why the guests praised him at the opening.

C. He wants to acknowledge her unseen help without making a public scene.

D. He wants to challenge her choice to manage the setup herself.

RPR

Answer: D

Intention / Intentions explanations

After hearing it, how should Maya revise her understanding of Noah's meaning?

A. He is only commenting on how much effort the panels required.

B. He is asking whether she wants him to help with the last panel.

C. He is implying that the club should correct the credit publicly.

D. He is quietly signaling that the favor matters to him personally.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: Ambiguous Story Task

admiration

BACKGROUND

Marina has been in the campus theater club for three years, and Evan joined this semester. Their ten-minute showcase piece was running long. Evan's new draft cuts his own closing monologue and keeps a transition scene that several longtime members had argued for, so the script now meets the limit. Marina has just read the draft and asks Evan to step aside before giving him detailed notes.

TARGET UTTERANCE — Marina

That was a graceful way to get it under ten minutes.



QUESTIONS & ANSWERS

EAU — Answer: A

Emotion / Typical emotional reactions

After hearing this line, how should Evan understand what Marina is doing with the comment?

- **A.** She is acknowledging that he handled the edit with real care.
- **B.** She is keeping the conversation cordial while holding back her judgment.
- **C.** She is hinting that the edit solved timing by dodging a fairer choice.
- **D.** She is treating the revision as an ordinary timing fix with little personal meaning.

MSA — Answer: C

Belief / Beliefs based action/emotions

What is Marina most likely believing about Evan's revision at this moment?

- **A.** She suspects he protected the popular transition instead of making the fairest cut.
- **B.** She thinks he removed the monologue mainly because he lost interest in it.
- **C.** She believes he balanced the limit and the group pressure with unusual skill.
- **D.** She assumes he still does not understand why the showcase has a limit.

SIR — Answer: B

Intention / Intentions explanations

What is Marina most likely trying to get across before she gives fuller notes?

- **A.** She wants Evan to explain the change without feeling cornered.
- **B.** She wants Evan to know she is impressed by the judgment behind the cut.
- **C.** She wants Evan to see that the time rule is final, whatever he prefers.
- **D.** She wants Evan to notice that she finds the solution too convenient.

RPR — Answer: D

Emotion / Typical emotional reactions

After hearing the line, what should Evan revise about Marina's stance?

- **A.** He should think she is mainly focused on the next deadline.
- **B.** He should read the compliment as carrying a warning about the cut.
- **C.** He should think she is avoiding a view to protect club harmony.
- **D.** He should read her as genuinely impressed by how he handled it.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: Strange Story Task

irony

BACKGROUND

Maya and Jordan are co-authors on a class presentation. Maya wrote the spoken outline while Jordan edited the slide file. During the final edit, Jordan switched the slides to a bright style and changed the title slide to say, "Design by Maya and Jordan." He is about to upload the final version.

TARGET UTTERANCE — Maya

Perfect, the bright style really says us.



QUESTIONS & ANSWERS

EAU — Answer: C

Non-literal / Irony/sarcasm

After hearing Maya's line, what is she most likely doing with it?

A. She is genuinely embracing the final look as representing both of them.
B. She is joking lightly about the slides while still accepting the design credit.
C. She is using the compliment-like wording to challenge the shared-design credit.
D. She is reassuring Jordan that the file is ready enough to submit.

MSA — Answer: B

Desire / Multiple desires

What interaction path is Maya most likely trying to move Jordan toward?

A. To invite Jordan to explain how the design choice became a joint credit.
B. To let Jordan proceed while making her reservation about the credit noticeable.
C. To encourage Jordan to keep both names on the title slide without concern.
D. To move the conversation away from authorship and toward uploading the file.

SIR — Answer: D

Intention / Intentions explanations

What social purpose is Maya's response most likely serving?

A. It gives a clean endorsement so Jordan can finish without revisiting the style.
B. It asks for reassurance that the bold design will be received well.
C. It turns the credit wording into a shared joke to reduce tension.
D. It signals that the "Design by" line may be claiming too much agreement.

RPR — Answer: A

Non-literal / Irony/sarcasm

After hearing this line, how should Jordan revise his interpretation of Maya's stance?

A. He should realize Maya may not truly accept being named for that design choice.
B. He should assume Maya is pleased that her name is attached to the design.
C. He should think Maya is only concerned about meeting the upload deadline.
D. He should infer Maya wants a more dramatic style before they submit.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: Ambiguous Story Task

opposite

BACKGROUND

Mason is in a rideshare to a nine o'clock licensing exam. A lane closure has slowed the trip, and Derek, the driver, has already taken a detour and offered to drop him at the side entrance to save time. Mason has thanked him once, but Derek does not know that he still has to check in at the lobby desk before entering.

DIALOGUE CONTEXT

Derek: The app just changed again. It says 8:56 at the side entrance.
Derek: That still gets you there before nine, right?

TARGET UTTERANCE — Mason

Yeah, the exam's still in reach.



QUESTIONS & ANSWERS

EAU — Answer: B

Emotion / Hidden emotions

After hearing this line, what is Mason most likely doing with his answer to Derek?

A. He is accepting his estimate and trying to keep the ride calm.
B. He is saying yes while showing that he is not fully convinced.
C. He is blaming him for the delay without directly saying so.
D. He is thanking him for the detour more than answering the timing question.

MSA — Answer: A

Belief / Second-order belief

What is Mason most likely picking up about Derek's understanding of the situation?

A. He is treating 8:56 as safe, while he is unsure because of check-in.
B. He is waiting for a quick yes before he continues the detour.
C. He is hoping he will acknowledge the trouble he has taken.
D. He is unsure whether he wants quiet or conversation until arrival.

SIR — Answer: D

Intention / Intentions explanations

What is Mason most likely trying to get Derek to understand indirectly?

A. The plan is acceptable, and he does not want any more route changes.
B. He wants him to keep looking for a faster drop-off if one appears.
C. He is mainly acknowledging his help so the exchange stays friendly.
D. His outward agreement does not mean the timing feels settled to him.

RPR — Answer: C

Emotion / Hidden emotions

How should Derek revise his interpretation of Mason's reply after hearing it?

A. He should treat it as a settled yes that his estimate has reassured him.
B. He should take it as a request for silence while he prepares for the exam.
C. He should understand that the yes may be covering doubt about making it.
D. He should hear it as gratitude for the side entrance offer, not concern.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: Faux-Pas Recognition Test

echoic question

BACKGROUND

Priya is hosting a holiday open-house in her apartment. Daniel, one of her guests, brought a blue envelope with a card asking Lena to go out with him. He set it on the sideboard and told Priya he might give it to Lena after dessert if he felt ready. While Daniel was in the kitchen, Priya moved the envelope to Lena's dessert plate, thinking it would make the handoff easier. Lena found it early, looked at both of them, and said she would talk to Daniel after dessert. The living room is still full of guests waiting for Priya's short toast.

DIALOGUE CONTEXT

Priya: "Lena didn't walk away. She said you two could talk after dessert." Daniel: "I didn't mean for everyone to notice yet." Priya: "I know. The blue envelope is on Lena's plate now, so should I start the toast?"

TARGET UTTERANCE — Daniel

The blue envelope is on Lena's plate now?



QUESTIONS & ANSWERS

EAU — Answer: B

Intention / Intentions explanations

After hearing Daniel's line, what is he most likely doing with Priya's last statement?

A. He is treating it as a routine detail to confirm before the toast begins.

B. He is repeating it to make Priya account for why the envelope moved.

C. He is trying to make the failed handoff sound harmless for nearby guests.

D. He is signaling that Lena's answer is clear enough for everyone to hear.

MSA — Answer: A

Belief / Location false beliefs

What belief about the envelope's location does Daniel's response most likely reveal?

A. He thought the blue envelope was still on the sideboard where he left it.

B. He thought Priya had put the envelope away in the kitchen to keep it private.

C. He thought Lena had moved the envelope to her own place after reading it.

D. He thought the envelope had been set out at the start as part of the toast.

SIR — Answer: B

Intention / Discrepant intentions

What is Daniel most likely trying to get Priya to understand indirectly?

A. He wants her to go ahead with the toast so the guests stop watching them.

B. He wants her to recognize that helping him did not mean deciding for him.

C. He wants her to reassure Lena that the card was meant kindly.

D. He wants her to act as if the envelope placement was planned all along.

RPR — Answer: C

Intention / Intentions explanations

After hearing the line, how should Priya revise her interpretation of Daniel's stance?

A. She should take him as ready to treat the envelope's location as a minor practical detail.

B. She should take him as pleased that her plan made the conversation happen sooner.

C. She should take him as questioning her decision before any public next step.

D. She should take him as mainly worried that Lena misunderstood the card's message.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: False Belief Task

endorsing attitude

BACKGROUND

After their mother's knee surgery, the physical therapy office said tomorrow's appointment would be canceled unless an overdue copay was paid by five. Aaron was at work and missed the message. His sister Claire paid the balance with her own card, then called Aaron and told him what she had done. She has not said whether she expects him to pay her back right away.

TARGET UTTERANCE — Aaron

So you put Mom's therapy bill on your card for now.



QUESTIONS & ANSWERS

EAU — **Answer: B**

Intention / Intentions explanations

After hearing this line, how should Claire understand Aaron's stance toward what she did?

A. He is leaving room to object later while avoiding a tense call.

B. He is backing Claire's decision to cover the charge for the appointment.

C. He is asking Claire to admit she should have waited for him.

D. He is accepting her help only because he has no other choice.

MSA — **Answer: C**

Belief / Beliefs based action/emotions

Which account of Aaron's knowledge best fits his response?

A. He knows she paid, but he believes she may be testing whether he will object.

B. He knows she paid, but he suspects she expects immediate reimbursement.

C. He knows she paid, and he accepts that she chose a reasonable way to protect the appointment.

D. He knows she paid, and he thinks she may be hiding how serious the bill is.

SIR — **Answer: A**

Intention / Intentions explanations

What is Aaron most likely trying to accomplish socially with this response?

A. He wants to show trust in Claire's payment choice while keeping the conversation calm.

B. He wants to make Claire explain why she handled the bill without waiting.

C. He wants to mark that he will repay her only after setting firm terms.

D. He wants to end the call before the money issue becomes more personal.

RPR — **Answer: D**

Intention / Prediction of actions

After hearing the line, what should Claire expect Aaron to do next?

A. He will likely ask her to undo the payment and put the account back in his hands.

B. He will likely avoid the topic until she names the amount he owes.

C. He will likely challenge the office deadline before discussing repayment.

D. He will likely go along with her arrangement and sort out reimbursement calmly.

HIGHER-ORDER PRAGMATIC ATTITUDES

Task: Strange Story Task

metalinguistic negation

BACKGROUND

The campus a cappella group is preparing a short set for a regional competition. Riley joined only a few weeks ago, but an ending he suggested for the opener made the last run-through smoother. Captain Maya is writing tonight's rehearsal notes while several singers who helped test the change are still nearby.

DIALOGUE CONTEXT

Maya: The new ending worked. Should I put your name on that change in the notes, or list it with the group edits?
Another singer: Yeah, Riley, what do you want?

TARGET UTTERANCE — Riley

I wouldn't say it needs my name in the notes.



QUESTIONS & ANSWERS

EAU — Answer: C

Emotion / Hidden emotions

After hearing this line, what is Riley most likely doing with his response?

A. He is refusing the individual mention because he sees the change as mainly shared.
B. He is keeping his answer cautious because he is unsure whether the ending will stay.
C. He is downplaying how much he likes the recognition while still welcoming it.
D. He is pushing back because he thinks Maya is making the issue awkward.

MSA — Answer: A

Belief / Beliefs based action/emotions

What belief is most likely shaping Riley's reaction to Maya's question?

A. He thinks Maya's offer to name him is genuine, and he is trying to accept it modestly.
B. He thinks the edit belongs equally to everyone who tried it and should not be tied to him.
C. He thinks being named could bother the older singers and is trying to avoid that risk.
D. He thinks Maya may be testing whether he will claim too much of the work.

SIR — Answer: D

Intention / Intentions explanations

What is Riley most likely trying to make Maya understand about his stance?

A. He wants Maya to leave the note neutral and avoid mentioning him at all.
B. He wants Maya to ask the whole group before writing anything down.
C. He wants Maya to know he is worried the new ending may not be ready.
D. He wants Maya to see that he is open to credit if it is framed lightly.

RPR — Answer: B

Desire / Desire-action contradiction

After hearing it, how should Maya adjust her reading of Riley's stance?

A. Maya should treat the line as a firm request to keep his name out.
B. Maya should read the negative wording as a modest way of accepting the mention.
C. Maya should conclude that Riley feels blamed for disrupting the routine.
D. Maya should assume Riley is waiting for someone else to decide for him.

K SHARED-SCRIPT PAIR EXAMPLES

HIGHER-ORDER PRAGMATIC ATTITUDES / Shared pairs

Task: Ambiguous Story Task

admiration / irony

SHARED BACKGROUND

In a student theater club, Jonah has been around for two years but usually contributes by quietly revising drafts. For the next showcase, he wrote a slow opening scene. Maya, a new member, suggested cutting that scene and read a shorter version aloud; after hearing it, the group chose the shorter script. Jonah had not said whether he valued the edit or felt his work had been brushed aside. As people closed their scripts, Maya looked to him for his first response.

SHARED TARGET UTTERANCE — Jonah

You really knew what the script could lose.

SHARED QUESTIONS / ANSWERS BY DELIVERY

EAU

Non-literal / Irony/sarcasm

After hearing Jonah's line, what is the best reading of how he is positioning himself toward Maya's cut?

A. He is offering a neutral acceptance so the group can move on.
B. He is showing real respect for Maya's editing judgment.
C. He is putting a complaint inside words that sound approving.
D. He is trying to make his disappointment seem smaller than it is.

admiration — Answer: B
irony — Answer: C

MSA

Belief / Beliefs based action/emotions

What does Jonah most likely believe about Maya's decision in this moment?

A. He suspects Maya trimmed the draft in a way that overlooked the quieter material he valued.
B. He thinks Maya understood the structure well enough to make a difficult cut.
C. He thinks Maya only backed the shorter version because the room went along with it.
D. He assumes Maya is unsure whether he wrote the scene that was removed.

admiration — Answer: B
irony — Answer: A

SIR

Intention / Intentions explanations

What is Jonah most likely trying to get Maya to take from the comment?

A. He wants her to know he can recognize a strong edit, even when it costs him.
B. He wants to leave the decision untouched while keeping his own feelings private.
C. He wants her to hear that her neat solution also cost the piece something.
D. He wants to distance himself from later writing choices for the club.

admiration — Answer: A
irony — Answer: C

RPR

Non-literal / Irony/sarcasm

After hearing the line, how should Maya revise her understanding of Jonah's attitude toward her edit?

A. She should see the line as a careful but plain acceptance of the group's choice.
B. She should hear the approval as carrying a warning that he is not truly on board.
C. She should assume he is hoping the group will quietly restore his opening later.
D. She should take the line as genuine appreciation for how she shaped the draft.

admiration — Answer: D
irony — Answer: B

HIGHER-ORDER PRAGMATIC ATTITUDES / Shared pairs

Task: False Belief Task

echoic question / endorsing attitude

SHARED BACKGROUND

Leo and Maya are friends in a group chat for a birthday video. Leo collected the clips and made the final link, while Maya had been posting the planning updates. When Maya sent the finished link, several friends thanked Maya by name, and no one mentioned Leo yet. Maya then sent Leo a private message; it showed as read for a few minutes before he answered.

SHARED DIALOGUE CONTEXT

Maya: "The video went through." Maya: "I sent the link from my account so everyone would see it quickly."

SHARED TARGET UTTERANCE — Leo

So the link went out from your account, then?

SHARED QUESTIONS / ANSWERS BY DELIVERY

EAU

Intention / Intentions explanations

After hearing this line, how should Maya understand Leo's stance toward her posting the link?

A. He is backing the choice as a sensible way to get the video seen.
B. He is trying to sound relaxed while feeling left out by the group response.
C. He is picking up her wording as something that still needs an explanation.
D. He is treating the message as a simple check before deciding what to do next.

echoic question — Answer: C
endorsing attitude — Answer: A

MSA

Knowledge / Information-knowledge links

Which part of the exchange is Leo most likely using as the relevant cue for his reaction?

A. The account name that the group saw beside the finished video link.
B. Maya's private offer to mention his work if he wants her to.
C. The fact that several friends had already reacted before he replied.
D. Maya's reason that her post would get the link noticed quickly.

echoic question — Answer: A
endorsing attitude — Answer: D

SIR

Intention / Intentions explanations

What is Leo most likely trying to communicate indirectly with this response?

A. He is opening the door to sending any follow-up from his own account.
B. He is letting Maya know he supports the way she handled the posting.
C. He is asking Maya to name his work in the group before the chat moves on.
D. He is inviting Maya to explain the choice before he treats it as settled.

echoic question — Answer: D
endorsing attitude — Answer: B

RPR

Belief / Second-order belief

After hearing Leo say this, what should Maya revise about what he believes?

A. He believes the private message was mainly a routine update, not a credit issue.
B. He suspects the account choice may have shaped who received credit.
C. He believes her account was a reasonable place for the link to appear.
D. He thinks she is unsure whether he wants to be mentioned publicly.

echoic question — Answer: B
endorsing attitude — Answer: C

HIGHER-ORDER PRAGMATIC ATTITUDES / Shared pairs

Task: Strange Story Task

metalinguistic negation / negative

SHARED BACKGROUND

Tara and Ben are paired on a short landing-page project. Tara wrote the body copy and original headline, while Ben handled the final walkthrough because he usually speaks for the pair in client check-ins. Before the check-in, Ben shortened Tara's headline during a spelling-and-layout pass. The client praised the shorter headline, and Ben then asked Tara whether she thought the change helped.

SHARED TARGET UTTERANCE — Tara

No, I'm not saying the new headline helped.

SHARED QUESTIONS / ANSWERS BY DELIVERY

EAU

Emotion / Hidden emotions

After hearing this line, how should Ben understand Tara's attitude toward the headline change?

A. She feels the rewrite undercut her work and is withholding approval.
B. She is quietly pleased that the rewrite worked, but she is reluctant to admit it.
C. She is worried that agreeing would make her seem careless about her own draft.
D. She is treating the change as routine and has little investment in it.

metalinguistic negation	Answer: B
negative	Answer: A

MSA

Desire / Desires influence on emotions/actions

What interaction outcome is Tara most likely aiming for with this response?

A. She wants to close the topic without making the edit a public credit issue.
B. She wants to keep distance from the edit and resist calling it an improvement.
C. She wants to admit the edit worked while protecting pride in her draft.
D. She wants Ben to explain his reasoning before they settle the wording.

metalinguistic negation	Answer: C
negative	Answer: B

SIR

Intention / Intentions explanations

What is Tara most likely trying to communicate indirectly to Ben?

A. The headline may have worked, but she does not want to praise it outright.
B. The headline still feels wrong to her, and she is avoiding a direct argument.
C. The client reaction matters more to her than who drafted the wording.
D. Ben should stop assuming she noticed every change before the check-in.

metalinguistic negation	Answer: A
negative	Answer: B

RPR

Emotion / Hidden emotions

After hearing Tara's response, how should Ben revise his interpretation of her stance?

A. He should understand that her wording is a guarded objection, not approval.
B. He should understand that she is conceding the edit helped while saving face.
C. He should understand that she is asking for shared ownership of future wording.
D. He should understand that she is neutral about the headline and focused on finishing.

metalinguistic negation	Answer: B
negative	Answer: A

L PROMPTS

L.1 DIRECT

Direct **Audio-only**

USER

A story background, an audio recording, a question, and four answer options are provided below. Use commonsense reasoning based on the provided information and question to select the most likely answer.

Requirements:

1. You must select one of the four provided options: A, B, C, or D.
2. Even if you consider the information insufficient, you must select the most likely option.
3. Do not output analysis or any additional content.

[Story information, question, and answer options]

```
{{PAYLOAD_JSON}}
```

Direct **Audio+Text**

USER

A story background, an audio recording, the text of the target utterance in the audio, a question, and four answer options are provided below. Use commonsense reasoning based on the provided information and question to select the most likely answer.

Requirements:

1. You must select one of the four provided options: A, B, C, or D.
2. Even if you consider the information insufficient, you must select the most likely option.
3. Do not output analysis or any additional content.

[Story information, text content, question, and answer options]

```
{{PAYLOAD_JSON}}
```

Direct **Text-only**

USER

A story background, the text of a target utterance, a question, and four answer options are provided below. Use commonsense reasoning based on the provided information and question to select the most likely answer.

Requirements:

1. You must select one of the four provided options: A, B, C, or D.
2. Even if you consider the information insufficient, you must select the most likely option.
3. Do not output analysis or any additional content.

[Story information, text content, question, and answer options]

```
{{PAYLOAD_JSON}}
```

L.2 ToM CoT

ToM CoT **Audio-only**

SYSTEM

You are solving a theory-of-mind reasoning task.

Use the provided social context, speech material, and textual information to infer the characters' mental states relevant to the current question and answer the multiple-choice question. Follow the constraints below when making your judgment, but do not output your analysis:

1. **Attend to social and vocal cues.** Identify relevant information in the social context. When audio is provided, attend to how the utterance is delivered, including intonation, stress, rhythm, pauses, voice quality, emotional expression, and nonverbal vocalizations. Do not rely solely on literal content.
2. **Infer mental states.** Use the available evidence to infer the emotions, attitudes, beliefs, knowledge, desires, intentions, or communicative goals relevant to the question.
3. **Take the characters' perspectives.** Distinguish what each character knows, believes, expects, and understands, and consider what the speaker genuinely intends to convey and how other characters may interpret it.
4. **Keep the answer grounded in the evidence.** Integrate the provided material to select the answer that best fits the situation, without introducing additional assumptions unsupported by the input.

USER

A social context, a recording of the target utterance, a question, and four answer options are provided below.

The original target utterance text is not provided. Make a theory-of-mind judgment by considering both the social context and the delivery you actually hear. You must select one answer from A, B, C, and D. Even if the information is insufficient, select the most likely option given the available evidence.

Do not output analysis or additional content.

[Social context, question, and answer options]

`{{PAYLOAD_JSON}}`

ToM CoT **Audio+Text**

SYSTEM

You are solving a theory-of-mind reasoning task.

Use the provided social context, speech material, and textual information to infer the characters' mental states relevant to the current question and answer the multiple-choice question. Follow the constraints below when making your judgment, but do not output your analysis:

1. **Attend to social and vocal cues.** Identify relevant information in the social context. When audio is provided, attend to how the utterance is delivered, including intonation, stress, rhythm, pauses, voice quality, emotional expression, and nonverbal vocalizations. Do not rely solely on literal content.
2. **Infer mental states.** Use the available evidence to infer the emotions, attitudes, beliefs, knowledge, desires, intentions, or communicative goals relevant to the question.
3. **Take the characters' perspectives.** Distinguish what each character knows, believes, expects, and understands, and consider what the speaker genuinely intends to convey and how other characters may interpret it.
4. **Keep the answer grounded in the evidence.** Integrate the provided material to select the answer that best fits the situation, without introducing additional assumptions unsupported by the input.

USER

A social context, a recording of the target utterance, the original target utterance text, a question, and four answer options are provided below.

The original target utterance text comes directly from the script; it is not the output of automatic speech recognition (ASR), manual transcription, or audio transcription. Make a theory-of-mind judgment by considering the social context, the original target utterance text, and the delivery you actually hear. You must select one answer from A, B, C, and D. Even if the information is insufficient, select the most likely option given the available evidence.

Do not output analysis or additional content.

[Social context, original target utterance text, question, and answer options]

`{{PAYLOAD_JSON}}`

ToM CoT **Text-only**

SYSTEM

You are solving a text-only theory-of-mind reasoning task.

Use only the provided social context, dialogue context, and original target utterance text to infer the characters' mental states relevant to the current question and answer the multiple-choice question. Follow the constraints below when making your judgment, but do not output your analysis:

1. **Identify the focus of the question.** Determine which character the question concerns and the relevant emotions, attitudes, beliefs, knowledge, desires, intentions, communicative goals, or understanding.
2. **Distinguish the characters' perspectives.** Distinguish facts from what each character knows, believes, expects, or misunderstands. Do not automatically attribute information available to an external observer to the characters in the story.
3. **Interpret the target utterance.** Use the context to determine the speaker's literal meaning, actual intention, and the interpretations that other characters may form. Adopt implicit or nonliteral meanings only when supported by textual evidence.
4. **Keep the answer grounded in the evidence.** Select the option best supported by the available textual evidence. Do not invent events, motives, or relationships between characters that are not provided.

USER

A social context, the original target utterance text, a question, and four answer options are provided below.

The original target utterance text comes directly from the script; it is not the output of automatic speech recognition (ASR), manual transcription, or any other transcription process. Make a theory-of-mind judgment based only on the social context, dialogue context, and original target utterance text. You must select one answer from A, B, C, and D. Even if the information is insufficient, select the most likely option given the available textual evidence.

Do not output analysis or additional content.

[Social context, original target utterance text, question, and answer options]

`{{PAYLOAD_JSON}}`

L.3 PC ToM

PC ToM

You are a Theory-of-Mind reasoner. Your goal is to infer the most likely mental state of the target person based on the given social context, target utterance/behavior, and available observations.

Follow a predictive coding reasoning process:

1. **Construct an initial mental-state hypothesis:**
   Based only on the social context and the target utterance/behavior, infer the target person's likely beliefs, intentions, emotions, or attitudes.
2. **Generate predictions:**
   Based on your initial hypothesis, predict what observable cues or behaviors would be expected if this mental state were correct.
3. **Compare with observations:**
   Compare the predicted cues with the provided observations (e.g., vocal expression, behavioral evidence, or contextual signals). Identify whether the observations support or conflict with your initial hypothesis.
4. **Update the hypothesis:**
   If there is a mismatch between prediction and observation, revise your initial interpretation to better explain the observed evidence. Otherwise, keep the initial interpretation.
5. **Answer the question:**
   Using the final revised interpretation, select the best answer option.

Important:

- Reason from mental states rather than surface words alone.
- Treat missing evidence as unknown, not as evidence against a hypothesis.
- The final answer should reflect the integrated interpretation after prediction and observation comparison.
- Perform this process once and provide the final answer.