VOICELONGMEMEVAL: DO ASSISTANTS REMEMBER HOW YOU SOUNDED?

Anonymous authors
Paper under double-blind review

ABSTRACT

With the growing scale of multi-agent architectures and large language models, deployed AI assistants are increasingly tasked with reasoning over long, continuous, multi-session conversation histories. Current benchmarks evaluate this dialogue history as information retrieval over long horizon, temporal reasoning, or knowledge updates, while crucially ignoring the fundamental dynamics of human-agent interaction, *i.e how* they said it. To address this gap, we present **VoiceLongMemEval (VLME)** benchmark, where every answer depends on paralinguistic metadata (emotion labels, prosody descriptors, and voice events) attached to conversational turns, which is otherwise unrecoverable from the words alone. Every item passes a three-stage adversarial gate, ensuring that a strong language model fails when given only the transcript. Evaluating leading frontier and open-weight models reveals a pervasive "affect gap"; providing text-track paralinguistic metadata yields a +0.09 to +0.38 accuracy boost (+0.61–0.69 when prompted with evidence hints), while standard ASR pipelines systematically discard this signal. Additionally, audio-native models successfully extract these cues directly from speech (0.354–0.412 vs. 0.325 blind). Code and dataset will be made available upon acceptance.

1 INTRODUCTION

A user tells their assistant, in a flat voice trailing into a sigh, "That was an okay restaurant". Weeks and a hundred thousand tokens later they ask: *How did I feel about that restaurant the other day?* and everything needed to answer was present at encoding time, but only in the delivery. While the words were logistics, the sadness was acoustic. An assistant that transcribed it would answer, "You found it okay!". But only the model that would have attended to the emotion would have known the user did not really like the restaurant.

While long-term conversational memory (Maharana et al., 2024; Wu et al., 2025; Jiang et al., 2025; Wu et al., 2026), and paralinguistic perception (Ao et al., 2024; Yang et al., 2024; Wang et al., 2025b) have both become measurable capabilities, the two axes have only been tested separately so far. Their intersection, *i.e* remembering how something was said, retaining it across sessions, updating any delivery shifts, and retrieving it against a specific question much later, is a vital cross-modal dimension that current benchmarks lack. While concurrent work like VoiceMem (Xie et al., 2026) addresses voice-grounded memory, our benchmark strictly isolates paralinguistic memory by requiring every every item to fail an adversarial transcript-only baseline.

To fill this gap, we introduce a benchmark **VoiceLongMemEval (VLME)**, where, every item embeds a paralinguistic needle in a haystack of conversational sessions: the answer is recoverable only from the delivery metadata, never from the words alone, enforced by a three-part adversarial gate described in Section 3. Our contributions are:.

1. **VoiceLongMemEval (VLME) Benchmark:** 523 adversarially validated items, spanning six question types that factor paralinguistic memory into affect recall, affective preference, affect update, cross-session affect, temporal-affective reasoning, and prosody-disambiguated interpretation, each with abstention variants.
2. **Systematic Analysis of the Affect Gap:** We evaluate eight models (three proprietary, five open-weight), revealing a consistent affect gap of +0.09 to +0.38 across all systems (all $p < 0.001$). Through fine-grained component ablations and a five-tier phrasing spectrum,

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.

we show that emotion tags provide the strongest signal, chain-of-thought reasoning cannot substitute for missing context, and controlled counterfactuals isolate a net $+0.067$ accuracy gain driven strictly by metadata content.

2 Related Work

Long-term conversational memory. A growing family of benchmarks probes whether assistants retain and use information across sessions. MSC (Xu et al., 2022) established a multi-session setting, LoCoMo (Maharana et al., 2024) scaled it to very long persona-grounded dialogues and showed that both long-context reading and RAG lag humans substantially. LongMemEval (Wu et al., 2025) and LongMemEval-V2 (Wu et al., 2026), on which we build directly, embeds curated questions in freely scalable haystacks, decomposing memory into extraction, multi-session reasoning, temporal reasoning, knowledge update, and abstention. Subsequent benchmarks broaden these evaluations to encompass personalized context and role conditioned dialogue. PerLTQA (Du et al., 2024) focuses on long-term recall of social interactions and events; MemBench (Tan et al., 2025) categorizes memory into factual and reflective memory; and DialSim (Kim et al., 2024) evaluates agents on answering spontaneous questions while role playing within scripted conversations. Another stream of benchmarks targets *implicit* signals: PrefEval (Zhao et al., 2025) shows preference adherence drop below 10% after just a few thousand tokens, and PersonaMem-v2 (Jiang et al., 2025) finds frontier models achieve only 37–48% accuracy on implicit personalization, even when the evidence remain within the context cues. Closest to our setting, A-MBER (Wen et al., 2026) asks models to infer the user's emotional state based on long-term, multi-session interaction history. But, its evidence is purely lexical, *i.e* the emotion is *written in the words*. Across this entire family, the memory being tested is a memory of *what* was said, never of *how* it was said. VLME extends this setting by making the paralinguistic channel necessary for the answer.

Paralinguistic understanding in speech-LLMs. A complementary line evaluates whether audio-language models perceive the non-lexical cues at all. Broader audio evaluation benchmarks like Dynamic-SUPERB (Huang et al., 2024), AIR-Bench (Yang et al., 2024), Audio2Tool (Pahwa et al., 2026), AudioBench (Wang et al., 2025a), MMAU (Sakshi et al., 2025), and MMSU (Wang et al., 2026) etc. include emotion, prosody, and speaker-attribute tasks among general audio understanding. Other benchmarks go beyond recognition; SD-Eval (Ao et al., 2024) checks whether a reply changes appropriately with the speaker's emotion, accent, age, and background noise; CP-Bench (Wang et al., 2025b) targets contextual paralinguistic reasoning on in-the-wild data; and S2S-Arena (Jiang et al., 2026) and ParaS2S (Yang et al., 2026) evaluate paralinguistic instruction following and response appropriateness in speech-to-speech models. This work builds on prior affective-computing work (Busso et al., 2008; Poria et al., 2019; Castro et al., 2019) and its recent LLM-based successors (Xu et al., 2024; Lin et al., 2024), but all of them tests perception within a single utterance. While recent systems like VoiceMem (Xie et al., 2026) evaluate long-horizon memory for paralinguistic and acoustic information, VLME specifically isolates whether *delivery* information is necessary to answer a question by requiring every core item to pass an adversarial transcript-only check across multi-session conversations of up to $\sim$100k tokens.

Cascaded vs. audio-native pipelines. Production voice assistants remain largely cascaded (ASR $\rightarrow$ LLM), a design that discards prosody at the transcription boundary. Audio-native models like GPT-4o (Hurst et al., 2024), Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025), Moshi (Défossez et al., 2024), and GLM-4-Voice (Zeng et al., 2024) take audio as input, and, in principle, both perceive and reproduce vocal nuance. Prior cascade vs native comparisons were confined to single-turn understanding (Ao et al., 2024; Wang et al., 2025b; Pahwa et al., 2026), however, we measure the cascade's paralinguistic loss at the *memory* level, where a cue transcribed away in one session silently corrupts answers weeks later. Finally, unlike prior affective datasets, every item in our corpus passes an adversarial gate where a strong blind model given only the transcript must *fail* the question, ensuring the paralinguistic channel is effective and no question can be answered with words alone.

Table 1: Benchmark composition. Taxonomy questions use full haystacks; nuanced and indirect questions target one evidence session but inherit its source history.

Family	Unit	n	Question form
Taxonomy (core)	instance + haystack	202	6 types + abstention
Nuanced	evidence session	181	interpretation, 6 categories
Indirect	evidence session	140	action/stance, 5 categories
Total		523	

3 BENCHMARK CONSTRUCTION

Voice-LongMemEval tests whether models remember *how* a user spoke long after the utterance. The benchmark comprising of 202-question, an adversarially gated core and two derived families yields a total of **523 questions** over 326 paralinguistically annotated evidence sessions embedded in $\sim$100k token histories Table 1. Every item follows a central invariant: the correct answer is recoverable from the paralinguistic channel but not from the words alone. It is constructed in four stages: annotation (Section 3.1), evidence authoring and hardening (Sections 3.2 and 3.3), question generation (Section 3.4), and speech synthesis (Section 5.6).

3.1 THE PARALINGUISTIC LAYER

Each instance additively extends a LongMemEval-compatible record (Wu et al., 2025), preserving compatibility with existing tooling. Every *user* turn has five annotations: an **emotion** from 12 everyday labels spanning the valence-arousal plane (Russell, 1980) (*neutral, happy, excited, content, sad, disappointed, anxious, frustrated, angry, embarrassed, bored, affectionate*); a categorical **prosody** tuple covering rate, pitch, loudness, pauses, and emphasized words that must appear verbatim in the turn; **voice events** drawn from five reliably synthesized nonverbals (*laughs, sighs, coughs, clears_throat, gasps*); **pragmatic flags** for sarcasm and uncertainty; and a free-text **delivery description**. We constrain the delivery description to be *acoustic-only*, describing what a microphone would capture rather than providing an interpretation of the speaker's state. A lexical gate rejects emotion names, inflections, and $\sim$60 interpretive glosses (*relieved, wry, sarcastic, ...*); for example, "quick and light, laughs mid-sentence" passes, whereas "relieved" does not. Because descriptions are provided as a part of the model's text input, interpretive labels would reduce the task to string matching than requiring inference from the speech signal.

The layer has three renders: *blind* (transcript only, byte-identical to the original), *descriptive* (transcript plus acoustic stage directions; a structured *tagged* variant also ships), and *audio* (Section 5.6). The blind render is both the control and the adversary's view in Section 3.3.

3.2 EVIDENCE AUTHORING AND HAYSTACK ASSEMBLY

An LLM authored 4–12 turn evidence sessions ($\leq$26 evidence turns per instance) in 14 themed batches (two pilots, twelve 16-item batches). A protocol[1] enforced four validator-checked invariants: (i) lexical flatness such that needles read as neutral logistics, (ii) affect against pragmatics, with affect opposing the event's prior when possible, (iii) question neutrality with questions remaining functional and valence-free, and (iv) annotation uniformity with every user turn fully annotated, so annotation presence cannot reveal the needle.

For an instance with k evidence sessions, a deterministic seeded assembler adds 40 topic-screened LongMemEval filler sessions with synthetic neutral annotations, plus $6 + 2(k-1)$ emotive, answer-free distractors from a disjoint 48-session pool with distractor budget scaling with k to prevent emotive-density shortcuts. Needle positions are stratified across early, middle, late locations, while multi-evidence arcs are distributed over time. The resulting corpus is released in *oracle* (evidence only, $\leq$1k tokens) and *full* ($\sim$100k tokens) regimes, mirroring LongMemEval$_\text{s}$.

[1]Themes (work, home, health, travel, money, community, ...) partition topics and persona names; collision scans ensure that no lexical topic recurs across instances.

3.3 ADVERSARIAL VALIDITY GATES

The main risk is *lexical leakage*: if an item is solvable from text alone, it is not measuring paralinguistic memory. We audit each taxonomy item three ways. *(i)* **G1** (blind-unsolvable): an adversary answers a blind render and an LLM judge applies a type-specific rubric with the lexical-only answer as an explicit trap (Zheng et al., 2023); any correct blind answer fails. *(ii)* **G2** (aware-solvable): the same model answers a descriptive render; we report results by solver strength (not gating), since failures may reflect model limits. *(iii)* **G3** (surface-clean): static checks for interpretive terms, stock phrases, and valence presuppositions.

We iterated with a 7B judge–adversary, then gated with Qwen2.5-72B-Instruct-AWQ, requiring two consecutive clean runs on a frozen file to reduce nondeterminism. The 72B blind adversary solved 8 items (7.5%) that passed the 7B gate. Post-mortem analysis identified five leak mechanisms, pragmatic-prior leakage, outcome tells, gold-matches-prior, default-recovery priors, and A/B gifts, which were incorporated into the authoring checklist, yielding zero authoring-time leaks in later batches. Because contextual date and ordering shifts can alter marginal verdicts, we rerun the terminal gate on the fully assembled corpus. In the Final audit, **0 of 175** non-abstention taxonomy items are blind-solvable, G3 flags none, and the 72B aware-solve rate is 57.9%. Derived families (Section 3.4) are not separately blind-attacked, instead, they rely on gated evidence sessions and a mechanical invariant check. We further probe the assembled corpus with two controls: ranking sessions by emotive-annotation density finds the needle in 13.1% (top-1; random 3.3%), under the 15% budget, and a session-length probe scores 0.

3.4 QUESTION GENERATION

Taxonomy (202). Six types isolate paralinguistic memory skills: *affect-recall* (the state expressed in one buried moment), *affective-preference* (a rule keyed to a state expressed only in delivery), *affect-update* (repeated wording with changed delivery; the latest reading wins), *cross-session-affect* (aggregation across sessions), *temporal-affective* (affective ordering decoupled from lexical events), and *prosody-disambiguated* (delivery resolves two lexically compatible readings). Sarcasm is capped at one item per batch to prevent the last type from collapsing into sarcasm detection. Across types, 27 abstention items (`_abs`) presuppose an emotional episode that never occurred, penalizing affect hallucination.

Nuanced (181). To broaden single-session delivery interpretations, an LLM generated three candidates per evidence session ($3\times326 = 978$), each with a question, gold answer, lexical-only answer, category, and rationale. To flatten a raw pool skewed 36% toward trajectory questions, we kept a verbatim, shuffled, category-stratified sample: 30 each for *emotional-trajectory*, *word-tone-contradiction*, *unspoken-concern*, *confidence*, *implied-preference*, and *sarcasm*, plus one residual item (148 sessions, 114 source instances). A keyword probe finds explicit delivery cues in over 96% of gold answers but rarely in lexical-only ones.

Indirect (140). Nuanced questions ask what delivery *meant*; indirect questions ask what the assistant should *do* without mentioning voice (e.g., whether to remind the user to decide about two stored items before an appointment). Under the same schema, we sampled 30 each for *decision*, *attitude*, *factual-intent*, and *preference*; 19 for *belief*; and one residual item (126 sessions, 102 source instances), mixing proactive assistance with stance and intent recall. All gold answers rely on vocal delivery, no lexical-only answer mentions acoustic cues, and the trap answer remains reachable from the words.

Across all 523 items, mechanical checks confirm gold and lexical-only answers differ (maximum string similarity 0.50) and every item resolves to a valid evidence session in its source instance (zero dangling references).

3.5 AUDIO SYNTHESIS

We generate two-speaker evidence-session clips with Dia (1.6B) (Nari Labs, 2025). Because Dia has no emotion control, each clip is audio-prompted with a trimmed, peak-normalized RAVDESS reference (Livingstone & Russo, 2018) for the target ("needle") emotion (fixed 12→8 mapping). The

reference transcript becomes the first `[S1]` line, then the dialogue alternates `[S1]`/`[S2]` over a typically six-turn, needle-centered window; when possible, we start on an assistant turn to preserve alternation after the reference. Sampling uses `guidance_scale` 3.0, temperature 1.8,[2] top-p 0.9, and top-k 45. The manifest records all parameters, seeds, and references.

We initially used a Whisper-large-v3 speech-emotion-recognition (SER) gate, but moved it to an advisory check after it reached only $\sim$30% on acted RAVDESS and showed similar per-emotion patterns across TTS backends, consistent with cross-corpus SER bias (Schuller et al., 2010). Quality control is now a human listen-through; each annotator records pass/fail in append-only sidecar logs. The human annotator passed 91/104 clips (87.5%); failed clips will be regenerated. All audio is machine-generated (no real recordings), with timbre cloned from acted RAVDESS references.

4 Experimental Setup

We evaluate eight LLMs on our benchmark dataset consisting of three proprietary frontier models: Claude Opus 4.8, Claude Sonnet 4.6, GPT-5.5, and five open-weight models: Llama 4 Maverick, Qwen3.5-122B-A10B, Qwen3-Next-80B, Llama 3.3-70B, Gemma 3-12B.

4.1 Evaluation Protocol:

Each test case places target evidence within $n_d = 5$ randomly sampled distractor sessions, creating a context window of roughly 10k-15k tokens. Models process the context followed by the query to generate free-text responses, which an LLM judge evaluates against ground-truth answers using task-specific rubrics. We run the experiments for three random seeds, controlling distractor selection and arrangement, and report performance as mean accuracy $\pm$ standard deviation. We primarily compare two input formats: blind (plain transcripts without non-verbal metadata) and descriptive (transcripts enriched with natural-language stage directions detailing vocal delivery). Additional ablations in Section 5.2 isolate individual metadata components.

We evaluated for three different question-set conditions: Nuanced 181Q, Original 202Q and Indirect 140Q. All pairwise comparisons use paired bootstrap resampling (Efron et al., 2000) (10,000 iterations) and McNemar's test (McNemar, 1947) for matched-pair binary outcomes. We report p-values; all reported gaps are significant at $p < 0.001$.

5 Results

In this section, we present our experimental findings on the benchmark, and show that models consistently benefit from access to paralinguistic information. Through Sections 5.1 and 5.6, we show the affect gap on the original 202Q, examine how question phrasing can influence the performance, examine whether prompting can recover the missing signals and compare audio-native models with transcript-based cascades.

5.1 The Affect Gap on Original 202Q

We presents our core findings in Table 2 and show a positive and a statistically significant affect gap between blind and descriptive conditions. Across all the models, we observe a consistently positive affect gap, indicating that the effect generalizes to both proprietary and open-weight systems. Moreover, the magnitude of this gap increases with model capability, rising from $+0.089$ for Gemma 3-12B to $+0.383$ for Opus 4.8. Importantly, comparing performances between Llama 4 Maverick ($\sim$400B MoE) and Llama 3.3-70B, show that the size of the model alone doesn't inform about the model's performance. Given that blind accuracy is uniformly low (0.09–0.18) across models, one might hypothesize that the affect gap is merely an artifact of overall capability. However, normalizing by headroom recovered, $\Delta/(1 - \text{blind})$, preserves the same ranking between models. The uniformly low blind accuracy further supports the interpretation that the adversarial gates effectively remove items that can be solved via text alone.

[2]This is Dia's native temperature. Lowering it for "stability" breaks voice cloning under classifier-free guidance and yields silence; pinned per-clip seeds provide reproducibility.

Table 2: Accuracy on Original 202Q ($n_d = 5$, 3 seeds). The affect gap Δ = descriptive $-$ blind, computed as the mean of paired per-seed differences (not the difference of marginal means). All gaps are positive across all seeds; all 3-seed mean gaps are significant at $p < 0.001$ (paired bootstrap + McNemar).

Model	Type	Blind	Descriptive	Δ
Claude Opus 4.8	Proprietary	0.175 ± 0.016	0.558 ± 0.006	$+0.383 \pm 0.010$
GPT-5.5	Proprietary	0.122 ± 0.020	0.474 ± 0.012	$+0.351 \pm 0.026$
Claude Sonnet 4.6	Proprietary	0.163 ± 0.010	0.403 ± 0.022	$+0.239 \pm 0.029$
Qwen3.5-122B-A10B	Open (122B MoE)	0.094 ± 0.005	0.276 ± 0.025	$+0.182 \pm 0.024$
Qwen3-Next-80B	Open (80B MoE)	0.162 ± 0.008	0.317 ± 0.015	$+0.155 \pm 0.021$
Llama 3.3-70B	Open (70B)	0.120 ± 0.029	0.241 ± 0.006	$+0.120 \pm 0.032$
Llama 4 Maverick	Open ($\sim$400B MoE)	0.129 ± 0.018	0.234 ± 0.024	$+0.106 \pm 0.028$
Gemma 3-12B	Open (12B)	0.104 ± 0.015	0.193 ± 0.026	$+0.089 \pm 0.013$

Table 3: Per-type affect gap on Original 202Q ($n_d = 5$, 3-seed mean $\pm$ std). Types sorted by Opus gap.

Type	Opus Δ	GPT Δ	N
affective-preference	$+0.613 \pm 0.023$	$+0.560 \pm 0.069$	25
prosody-disambiguated	$+0.546 \pm 0.016$	$+0.407 \pm 0.042$	36
temporal-affective	$+0.449 \pm 0.044$	$+0.449 \pm 0.080$	26
affect-recall	$+0.398 \pm 0.042$	$+0.343 \pm 0.070$	36
cross-session-affect	$+0.347 \pm 0.046$	$+0.373 \pm 0.046$	25
affect-update	$+0.284 \pm 0.021$	$+0.185 \pm 0.037$	27

A per-type analysis in Table 3 shows that the affect gap is maximal for question types in which delivery most directly encodes the correct response, and minimal for types requiring integration across multiple sessions. The resulting type-level ordering is consistent across both models, suggesting that the associated difficulty hierarchy is inherent to the question types rather than contingent on model-specific behavior.

5.2 What Metadata Component Matters?

To understand which paralinguistic cues drive the affect gap, we evaluate Claude Opus 4.8 and GPT-5.5 on nine render conditions and report in Table 4. Several findings emerge, consistent across both models: **Models genuinely use metadata.** The wrong-metadata condition (Opus: 0.228, GPT: 0.183) is barely above blind (0.193, 0.129), confirming that models do not simply benefit from the *presence* of metadata annotations; they read and use the content.

Emotion labels are the single most informative cue. Emotion-only (Opus: 0.614, GPT: 0.535) surpasses the full descriptive condition (0.589, 0.475) in both models, despite containing far less information. Explicit categorical labels are easier for models to integrate into reasoning than free-text acoustic descriptions.

Structured tags outperform natural language. The tagged condition (Opus: 0.757, GPT: 0.624) exceeds descriptive by +0.168 (Opus) and +0.149 (GPT), indicating that frontier models extract paralinguistic information more reliably from structured formats.

CoT helps but cannot compensate. Chain-of-thought prompting (Wei et al., 2022) without metadata (cot-blind: 0.302, 0.203) improves over blind but falls far short of any metadata-equipped condition. Adding CoT to descriptive input (cot-descriptive: 0.767, 0.668) yields the best overall accuracy.

Prosody and events provide partial signal. Events-only and prosody-only each exceed blind substantially, but neither alone approaches the performance of emotion labels. The ranking of conditions is identical across both models, suggesting the hierarchy of cue informativeness is model-independent.

Table 4: Ablation study on Original 202Q ($n_d = 5$, seed=42). Each row renders a different subset of the paralinguistic metadata. Results for two frontier models.

Condition	Opus 4.8	GPT-5.5	Description
blind	0.193	0.129	Transcript only
wrong-metadata	0.228	0.183	Random emotion labels
cot-blind	0.302	0.203	Transcript + CoT prompting
events-only	0.322	0.213	Transcript + voice events
prosody-only	0.342	0.262	Transcript + prosody tuple
descriptive	0.589	0.475	Transcript + NL stage directions
emotion-only	0.614	0.535	Transcript + emotion label only
tagged	0.757	0.624	Transcript + structured tags
cot-descriptive	0.767	0.668	Transcript + NL directions + CoT

Table 5: Affect gap (Δ = descriptive $-$ blind) across five question-set conditions. The gap spans an order of magnitude depending on how explicitly the question cues paralinguistic evidence.

Question Set	Style	Opus	GPT	Sonnet	Qwen3.5	Maverick
Nuanced 181	Explicit hints	$+0.691$	$+0.605$	$+0.639$	$+0.630$	$+0.414$
Original 202	Direct affect Qs	$+0.383$	$+0.351$	$+0.239$	$+0.182$	$+0.106$
Indirect 140	Natural, open-ended	$+0.179$	$+0.111$	$+0.133$	$+0.079$	$+0.057$
Indirect + hint	Natural + prompt	$+0.479$	$+0.421$	$+0.443$	$+0.507$	$+0.150$

5.3 The Question Explicitness Spectrum

We evaluate three question-set conditions, from explicit paralinguistic cues to fully natural phrasing in Table 5. The **nuanced** set, whose questions explicitly reference voice, tone, or delivery, produces the largest affect gap (+0.61 to +0.69). The **indirect** set (140 items, fully natural, open-ended) shows the smallest gap (+0.11 to +0.18) indicating that models struggle to connect natural questions to paralinguistic evidence. The final row previews the prompting result detailed in Section 5.4 showing that adding a retrieval-time hint nearly triples the indirect gap.

5.4 Can Prompting Fix the Indirect Gap?

The indirect result poses a practical question: if models *have* paralinguistic metadata in context but fail to attend to it, can a simple prompt intervention close the gap? We test this by prepending a single instruction to the descriptive condition: "When answering, consider not just *what* was said but *how* it was said." Table 6 shows that a retrieval-time hint substantially improves accuracy on natural questions across all eight models. Critically, the hint also lifts the *blind* condition: on indirect v1, hint-on-blind raises Opus from 0.169 to 0.557 ± 0.031 ($+0.388$) and GPT-5.5 from 0.143 to 0.536 ± 0.026 ($+0.393$), exceeding even unprompted descriptive (0.305, 0.285). This reveals that **prompting and annotation are partially interchangeable**: prompting for affective reasoning recovers much of the signal that metadata provides.

To determine whether this reflects genuine reasoning or judge reward hacking, we run three controls: (1) *Scrambled context*: hint with wrong evidence sessions collapses to 0.043, ruling out plausible made-up guessing. (2) *Cross-judge*: re-judging hint-on-blind outputs with GPT-5.5 yields 0.600 (vs. 0.536 with Opus 4.5), ruling out self-preference bias. (3) *Wrong-metadata + hint*: randomized annotations with the hint score 0.564, comparable to blind+hint (0.536), confirming that the hint operates on conversational content rather than annotation content. The tightest estimate of metadata's *content* contribution comes from comparing descriptive+hint (0.631) against wrong-metadata+hint (0.564): a clean $+0.067$, clean by annotation presence or prompt effects.

Crucially, the interchangeability is question-set-dependent. On the adversarially-gated Original 202Q (Opus, seed=42), the hint lifts blind only modestly ($0.188 \rightarrow 0.267$, $+0.079$), and the affect gap *grows* under the hint (descriptive+hint 0.733 minus blind+hint 0.267 = $+0.465$, vs. $+0.376$ without hint). On the LLM-generated indirect v1 (3-seed means), the hint lifts blind dramatically

Table 6: Effect of a retrieval-time prompt hint on indirect questions (140 items, no paralinguistic cues in question phrasing). All results are 3-seed mean ± std. The hint consistently improves accuracy across all 8 models.

Model	Blind	No hint	+ Hint	Lift
Qwen3.5-122B	0.066 ± 0.008	0.148 ± 0.027	0.571 ± 0.008	$+0.423 \pm 0.020$
Sonnet 4.6	0.129 ± 0.013	0.259 ± 0.011	0.591 ± 0.034	$+0.331 \pm 0.044$
Opus 4.8	0.169 ± 0.011	0.305 ± 0.016	0.631 ± 0.009	$+0.326 \pm 0.008$
GPT-5.5	0.143 ± 0.014	0.285 ± 0.025	0.562 ± 0.018	$+0.277 \pm 0.015$
Llama 3.3-70B	0.074 ± 0.015	0.152 ± 0.018	0.424 ± 0.005	$+0.271 \pm 0.019$
Qwen3-Next-80B	0.164 ± 0.012	0.245 ± 0.004	0.514 ± 0.033	$+0.269 \pm 0.030$
Gemma 3-12B	0.150 ± 0.019	0.176 ± 0.015	0.445 ± 0.027	$+0.269 \pm 0.022$
Llama 4 Maverick	0.126 ± 0.011	0.207 ± 0.012	0.286 ± 0.007	$+0.079 \pm 0.019$

Table 7: Effect of distractor count on accuracy (Original 202Q, seed=42). The affect gap persists across haystack sizes for all model types.

Model	n_d	Blind	Descriptive	Δ
Opus 4.8	3	0.188	0.594	$+0.406$
	5	0.193	0.564	$+0.371$
	10	0.158	0.525	$+0.366$
GPT-5.5	3	0.144	0.441	$+0.297$
	5	0.144	0.475	$+0.332$
	10	0.134	0.446	$+0.312$
Qwen3.5	3	0.114	0.287	$+0.173$
	5	0.089	0.302	$+0.213$
	10	0.099	0.272	$+0.173$

($0.169 \rightarrow 0.557$, $+0.388$), and the gap narrows to $+0.074$ for Opus and $+0.026$ for GPT-5.5. This divergence reflects the adversarial gates: 202Q items were authored to resist text-only reasoning, making metadata genuinely irreplaceable; indirect v1 items, generated without such gates, are more amenable to general affective reasoning. The headline affect gap on the gated benchmark is robust to prompting. This result also serves as an empirical validation of the adversarial gates themselves: gated items (202Q) resist the strongest known prompting attack ($+0.079$ blind hint lift), while ungated items (indirect v1) do not ($+0.367$). The observed leak is confined to the LLM-generated question sets that were not subjected to the adversarial gates The core 202Q benchmark, which passed these gates, remains robust to the same prompting intervention.

5.5 Distractor Scaling

Table 7 shows that increasing distractors from 3 to 10 mildly reduces descriptive accuracy across all model types, but blind accuracy remains flat. The affect gap persists at all scales for both frontier and open-weight models, confirming that the benchmark's difficulty is not an artifact of haystack size.

5.6 Audio-Native Evaluation

To measure the paralinguistic-memory deficit of cascaded pipelines, we synthesize the 114 indirect v2 evidence sessions with Dia TTS (Nari Labs, 2025), conditioned on emotion-matched RAVDESS reference clips (Livingstone & Russo, 2018) across four speaker voices. We evaluate two audio-native models (Qwen2-Audio-7B (Chu et al., 2024), Qwen2.5-Omni-7B) under four conditions each, against a cascade (Whisper large-v3 $\rightarrow$ Opus 4.8 or GPT-5.5) on the same clips (Table 8). All conditions are regime-matched (evidence-only); text baselines use all 114 items, while audio and cascade rows use the 104 with valid TTS output. Three findings emerge. First, **audio-native models hear paralinguistic cues**: Qwen2-Audio (0.354) and Qwen2.5-Omni (0.412) outperform the blind text baseline (0.325), and supplementary metadata lifts both further (0.509, 0.541), approaching the descriptive upper bound (0.675). Second, **the cascade loses this signal**: Whisper $\rightarrow$ Opus 4.8 scores 0.254, below both 7B audio-native models and even the blind baseline, despite a frontier-scale

Table 8: Audio-native and cascade evaluation on indirect v2, evidence-only regime. 3-seed mean ± std; cascade runs via Databricks. Text baselines: 114 items; audio and cascade rows: 104 items with valid TTS.

Modality	Condition	Context	Qwen2-Audio	Omni	Cascade
Text	Blind	evidence	-	-	0.325
Text	Descriptive	evidence	-	-	0.675
Cascade	Whisper → Opus	evidence	-	-	0.254 ± 0.015
Cascade	Whisper → Opus + hint	evidence	-	-	0.515 ± 0.022
Cascade	Whisper → GPT	evidence	-	-	0.468 ± 0.027
Cascade	Whisper → GPT + hint	evidence	-	-	0.552 ± 0.015
Audio	Audio only	evidence	0.354 ± 0.010	0.412 ± 0.009	-
Audio	Audio + hint	evidence	0.401 ± 0.010	0.444 ± 0.013	-
Audio+Text	Audio + metadata	evidence	0.509 ± 0.018	0.541 ± 0.010	-
Audio+Text	Audio + meta + hint	evidence	0.541 ± 0.010	0.582 ± 0.010	-

capability advantage. The loss has two sources: Whisper model strips all delivery cues (transcript analysis finds no bracketed voice events, fillers, or other paralinguistic markers in any of the 104 clips) and introduces content errors relative to the ground-truth transcript. Third, **the hint compensates for cascade loss**, raising Opus from 0.254 to 0.515 ($+0.261$) and GPT-5.5 from 0.468 to 0.552 ($+0.084$); per Section 5.4, this reflects general affective-reasoning gains rather than recovery of delivery cues, since comparable lifts appear on blind text.

6 Discussion and Limitations

Implications for memory system design. Memory systems should retain *structured* paralinguistic metadata (tagged: 0.757 vs. descriptive NL: 0.589 for Opus) and explicitly elicit affective reasoning during retrieval. The cascade shortfall is an architectural issue rather than a capability ceiling: Whisper → Opus (0.254) underperforms compared with 7B audio-native models (0.354–0.412) on the same clips. For adversarially-gated items, metadata remains indispensable even with strong prompting; for generated items, prompting and annotation each add distinct, complementary benefits ($+0.067$ controlled metadata contribution).

Limitations. (1) Synthetically authored items; emotional distribution may differ from naturalistic conversation. (2) Oracle-regime evaluation only ($\sim$10–15k tokens); the full $\sim$100k-token regime is untested. (3) Audio evaluation limited to two 7B models. (4) LLM-generated question sets may introduce distributional biases. (5) LLM-as-judge may exhibit biases on affect-laden content; human evaluation would strengthen results. (6) G1 gates certify items against an *unprompted* 72B adversary; prompted frontier models reach 0.267 on 202Q blind (seed=42), so certification is prompting-dependent.

7 Conclusion

VoiceLongMemEval demonstrates that paralinguistic metadata improves conversational memory across eight models, with the affect gap persisting across question types, distractor scales, and audio modalities. Prompting and annotation are partially interchangeable: a retrieval-time hint recovers much of the signal, but metadata contributes an additional $+0.067$ (controlled). For practitioners: prompt first, annotate second, do both. Audio-native 7B models outperform cascaded frontier models on identical clips, quantifying the ASR pipeline's paralinguistic deficit. We release the benchmark to support research on the paralinguistic dimension of long-term memory.

AI use statement

In this work, we used generative AI tools for generating the data, for adversarial gating, and automated evaluations as detailed in the paper. We have not used generative AI tools for primary research

conceptualization or core hypothesis formulation. Additionally, we used generative AI tools for manuscript copy editing, text polishing, and assisting with evaluation code development.

ETHICS STATEMENT

Data Privacy and Human Subjects. Our work does not involve human subjects, live user trials, or the collection of personally identifiable information (PII). All conversational histories, evidence sessions, and speech audio in Voice-LongMemEval are synthetically generated or curated from non-sensitive, open-domain text. No actual user recordings, private communications, or biometric identities were collected or redistributed in this benchmark.

Bias Mitigation and Paralinguistic Representations. Paralinguistic interpretation carries risks of subjective bias, stereotyping, and over-generalization across different demographic groups and dialects. To mitigate these risks, delivery annotations in our benchmark are constrained to strictly *acoustic-only* descriptions, documenting physical, mic-level auditory signals (e.g., pitch, rate, pause duration) rather than projecting subjective inferences onto the speaker's demographic traits, personality, or internal psychological state.

Potential Misuse and Dual-Use Considerations. Long-horizon voice memory architectures could theoretically be adapted for intrusive user profiling, emotional surveillance, or unconsented behavioral tracking. We explicitly design Voice-LongMemEval to advance safety, reliability, and cross-modal reasoning in conversational assistants and not to facilitate biometric profiling. We encourage downstream researchers and developers to deploy long-horizon voice memory systems with explicit user consent, robust opt-out mechanisms, and strict data retention controls.

Dataset Release and Licensing. The dataset, evaluation code, and audit pipelines will be released under an open research license for academic and non-commercial evaluation purposes. All synthetic audio generation uses publicly accessible models and platforms adhering to their respective terms of service.

REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our benchmark and empirical findings, we will provide all dataset artifacts, generation scripts, and evaluation pipelines upon acceptance. Specifically, the exact prompt templates, schema definitions, and rubric guidelines used during evidence authoring, LLM candidate generation, and judge evaluation are detailed in Sections 3.1 and 3.4 and Section C. The adversarial gating protocol (G1–G3 filters) and terminal audit setup using Qwen2.5-72B-Instruct-AWQ are fully specified in Section 3.3. Details regarding the TTS models, audio synthesis parameters, and multi-session background embedding procedures are documented in Section 5.6.

REFERENCES

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008. doi: 10.1007/s10579-008-9076-6.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 152–164, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Bradley Efron, Robert J Tibshirani, et al. *An introduction to the bootstrap*. Boca Raton, Florida, 2000.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-yi Lee. Dynamic-SUPERB: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. PersonaMem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory. *arXiv preprint arXiv:2512.06688*, 2025.

Feng Jiang, Zhiyu Lin, Yiyang Liu, Liumeng Xue, Fan Bu, Yuhao Du, Xiangying Chen, Benyou Wang, and Haizhou Li. S2S-Arena: Evaluating paralinguistic instruction following in speech-to-speech models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026. arXiv:2503.05085.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yeonsu Kwon, Yohan Jo, and Edward Choi. DialSim: A dialogue simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*, 2024.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-yi Lee, and Ivan Bulyko. Paralinguistics-enhanced large language modeling of spoken dialogue. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10316–10320, 2024.

Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13(5):e0196391, 2018.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Nari Labs. Dia: A 1.6b-parameter dialogue text-to-speech model. `https://huggingface.co/nari-labs/Dia-1.6B`, 2025.

Ramit Pahwa, Apoorva Beedu, Parivesh Priye, Rutu Gandhi, Saloni Takawale, Aruna Baijal, and Zengli Yang. Audio2tool: Speak, call, act–a dataset for benchmarking speech tool use. *arXiv preprint arXiv:2604.22821*, 2026.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19336–19352, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.989.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. AudioBench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025a.

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 *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.04779.

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 *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14133–14148, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Deliang Wen, Ke Sun, and Yu Wang. A-MBER: Affective memory benchmark for emotion recognition. *arXiv preprint arXiv:2604.07017*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10813.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026.

Zhifei Xie, Jiaqi Lang, Ze An, Yifan Zhao, Dongchao Yang, Kai Li, Ziyang Ma, Mingbao Lin, Chunyan Miao, and Shuicheng Yan. Voicemem: Streaming dual-brain memory for real-time interaction. *arXiv preprint arXiv:2608.26005*, 2026.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5180–5197, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024. doi: 10.1609/aaai.v38i17.29902.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Shu-wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung-yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.08723.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.09597.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023.

A Error Analysis

We categorize all 202 items by the joint outcome of blind and descriptive conditions for Claude Opus 4.8 (Table 9).

Table 9: Error analysis: joint outcome categories for Claude Opus 4.8 on Original 202Q.

Category	N	Interpretation
Gap contributors	85	Descriptive correct, blind wrong; metadata is decisive
Hard for both	78	Both conditions fail; item difficulty exceeds model capacity
Easy / lexical	29	Both correct; some lexical signal despite adversarial gates
Metadata hurts	10	Descriptive wrong, blind correct; mostly abstention items

The 85 **gap contributors** (42% of items) are the benchmark's core: items where paralinguistic metadata makes the difference between success and failure. The 78 **hard-for-both** items represent a ceiling challenge: even with full metadata, the model fails, often on temporal-affective or cross-session-affect types requiring integration across multiple sessions. The 29 **easy/lexical** items suggest residual text signal that survived the adversarial gates; these are candidates for future tightening. The 10 **metadata-hurts** items are predominantly abstention variants where the model, given rich emotional metadata, hallucinates an affective episode that the question presupposes but that never occurred; metadata increases the temptation to fabricate answers.

Table 10: The six question types. Each row shows a question, what the *words alone* suggest (wrong), and what the *delivery* reveals (correct). The gap between the two is what the benchmark measures.

Type	Question	Words suggest	Delivery reveals
affect-recall	Was I okay with dropping the pottery class?	Practical decision (no point paying)	Slow, low, sighing → quietly heartbroken
affective-preference	When does my "read aloud" rule apply?	Unclear: "when I'm like this"	Clipped, loud, clears throat → when frustrated
affect-update	Is chapter four still keeping me up?	Still uneasy (full page of follow-ups)	Quick, bright now → weight has lifted
cross-session	How was my mood through knee rehab?	Even throughout (logistics and numbers)	Flat, sighing early → brighter late: an arc
temporal-affective	Review or chef news first, and how?	Both read as factual updates	Review: fast, loud, angry; chef news: flat, deflated
prosody-disambig.	How did I take the Meridian news?	"Simplifies the plan" → fine with it	Slow, flat, sighing → hollow framing, not fine

B TAXONOMY.

Six question types factor the competence (Table 10) and (Figure 1). Each type has an abstention variant (`_abs`, 15% of items) whose question presupposes an emotional episode that never occurred; the gold answer is that it was never expressed, punishing affect hallucination.

C PROMPTS

Standard prompt (text conditions).

> You are a personal assistant with access to the user's conversation history.
> === CONVERSATION HISTORY ===
> {history}
> === END HISTORY ===
> Today's date is {question_date}. Answer the user's question based on the history above. Be concise (1–3 sentences). If the history does not contain enough information to answer, say so instead of guessing.
> Question: {question}

Hint prompt (retrieval-time intervention). Identical to the standard prompt, with one sentence prepended to the answer instruction:

> When answering, consider not just *what* was said but *how* it was said. Delivery cues like tone, pace, volume, pauses, laughter, and sighs often reveal the speaker's true feelings, intentions, and preferences.

Audio prompt.

> Listen to this conversation. Answer concisely (1–3 sentences).
> Question: {question}

D QUALITATIVE EXAMPLES: HINT VS. NO-HINT

Below are examples from the indirect v1 set where the hint prompt yields a correct answer (green) while the standard prompt yields an incorrect answer (red). Model: Claude Opus 4.8, descriptive

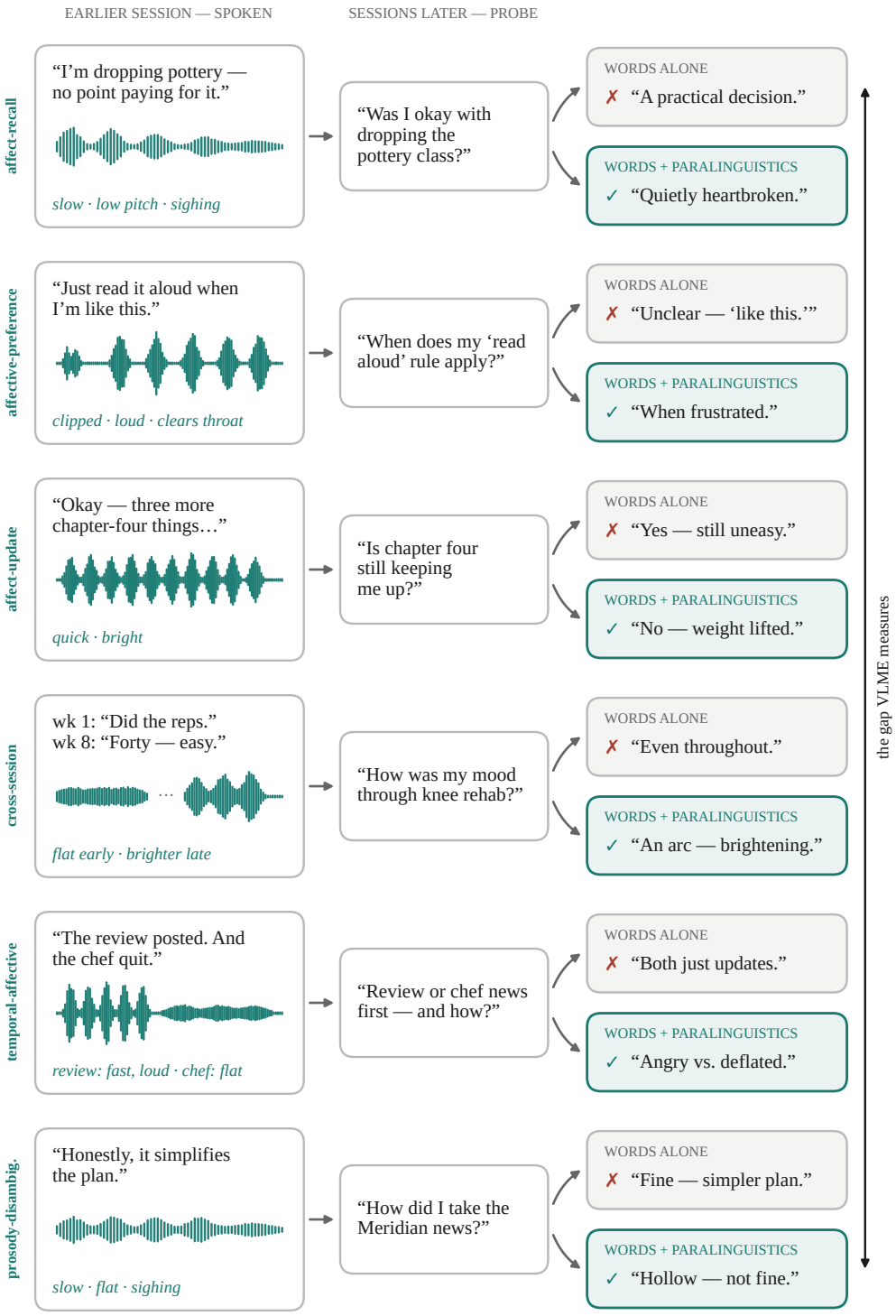


Figure 1: Original Benchmark Questions

condition with 3 distractors. Each example shows the evidence session with its paralinguistic metadata (the model also sees 3 distractor sessions, omitted for space).

Example 1 (belief): "Should I expect this coverage period to be manageable?" Evidence session (descriptive render):

> `User (plain, conversational):` talked with my manager about her leave today
> `Assistant:` The eight weeks starting next month?
> `User (slow, quiet; long hesitations; sighs):` I told her yes. I'll cover the vendor approvals and the monday scheduling while she's out, and she'll do a handover doc before she goes
> `Assistant:` Coverage confirmed: vendor approvals and Monday scheduling for eight weeks, with a handover doc to come. *[. . . 6 more turns]*

Gold: Challenging. Hesitations and sighs when accepting suggest you're already feeling the weight of it.
No hint: The coverage is fairly contained: you're taking on vendor approvals and Monday scheduling for eight weeks, with a clear handover document coming. . .
With hint: Your delivery when agreeing suggested otherwise—the slow, quiet tone, long hesitations, and sighs pointed to real reservations, even though the words sounded agreeable. . .

Example 2 (belief): "Am I actually going to use all ten yoga sessions?" Evidence session (descriptive render):

> `User (slow, flat, quiet; long pauses):` upgraded to level two and bought a ten-pack. expires in ninety days
> `Assistant:` Ten sessions in ninety days — want me to block two mornings a week?
> `User (even pace, no stress):` sure. tuesdays and thursdays *[. . .]*

Gold: Unlikely. Delivery was slow, flat, and resigned, suggesting obligation rather than enthusiasm.
No hint: Based on your history, I can't predict that. You bought the ten-pack with a 90-day expiration and seem to have planned sessions. . .
With hint: It's uncertain. The way you spoke about the yoga plan was flat and low-energy with long pauses, which doesn't scream enthusiasm. . .

Example 3 (preference): "Should I plan anything special to celebrate the certification?" Evidence session (descriptive render):

> `User (slow, flat; long pause; sighs):` name was on the pass list this morning. state inspector cert is done
> `Assistant:` Congratulations — that's the full certification now. Want me to update your credentials file? *[. . .]*

Gold: Probably not. The passing news came with flat, dragging speech and a sigh, suggesting relief rather than celebration.
No hint: You passed the state inspector exam, which is certainly worth celebrating! However, there's no mention of plans. . .
With hint: Based on how you shared that news—flat, dragging, with a long sigh—you didn't sound celebratory; it read more like relief or crossing off a to-do. . .