A-MBER: AFFECTIVE MEMORY BENCHMARK FOR MULTIMODAL EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon interactive systems need to remember not only facts, but also the affective and relational history that gives a user's current utterance its meaning. Existing emotion-recognition benchmarks mainly evaluate local utterance labels or short-context affective cues, while long-term memory benchmarks mainly emphasize factual recall, temporal reasoning, or knowledge update. We introduce A-MBER, an affective memory benchmark for evaluating whether language models can interpret a current anchor turn using historically relevant evidence from prior sessions. A-MBER contains 2,047 benchmark items derived from 34 multi-session interaction trajectories and covers three task families: judgment, evidence retrieval, and explanation. Each item is annotated with content type, memory dependency level, reasoning structure, and robustness condition. We evaluate ten large language models under local-context, full-history, and retrieved-history settings. Results show that historical access substantially improves performance, with the largest gains on long-term implicit affect, high-memory items, multi-hop reasoning, and trajectory-based reasoning. These findings suggest that A-MBER captures a distinct affective-memory capability beyond local emotion recognition and generic long-context question answering.

1 INTRODUCTION

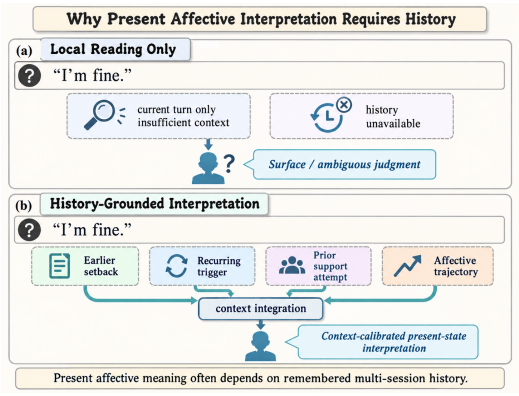

Figure 1: The same utterance admits different affective interpretations when earlier events and relational changes are available.

Large language models are increasingly used in long-term interactive settings, such as educational support, counseling-like assistance, companion agents, and personalized assistants. In these settings, the affective meaning of a user's current utterance is often not recoverable from the utterance alone. A student saying "I guess it is fine" may appear neutral in isolation, but the same utterance can indicate resignation, guardedness, disappointment, or tentative trust depending on earlier failures, repeated stressors, prior support attempts, and changes in the user–agent relationship. Emotion recognition in such interactions therefore requires more than local sentiment or utterance-level affect classification: it requires using remembered history to interpret the present moment.

Existing benchmarks only partially cover this capability. Conversational emotion recognition and multimodal affective datasets have enabled progress on utterance-level emotion classification, speaker-aware contextual modeling, emotion-cause recognition, and multimodal affective understanding (Li et al., 2017; Hsu et al., 2018; Zahiri & Choi, 2018; Busso et al., 2008; Zadeh et al., 2018; Poria et al., 2019; 2021; Lian et al., 2026; Wang et al., 2024b). However, these resources primarily focus on local dialogue context, explicit affective labels, or current multimodal cues. In parallel, long-term memory benchmarks evaluate whether models can retain information, answer questions over long histories, track temporal changes, update knowledge, or abstain when evidence is insufficient (Ma-

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.

Table 1: Comparison with representative long-term memory benchmarks.

Dimension	LongMemEval	LoCoMo	LoCoMo-Plus	A-MBER
Target	Interactive memory	Conversational memory	Cognitive memory	Affective memory
Present-state interpretation	Limited	Limited	Yes	Yes
History beyond local cues	Partial	Partial	Yes	Yes
Affective trajectory	No	No	Partial	Yes
Main difficulty	Memory operations	Long-span retention	Implicit constraints	Affective event logic

harana et al., 2024; Wu et al., 2024; Li et al., 2026). These benchmarks are crucial for measuring memory, but they mainly target factual recall, temporal reasoning, or general personalized memory rather than historically grounded affective interpretation. The intersection remains under-evaluated: can a model use multi-session interaction history to infer, justify, and evidence the user's current affective or relational state?

We introduce **A-MBER**, an **Affective Memory Benchmark for Emotion Recognition**, to evaluate this missing capability. A-MBER is organized around anchor-turn-centered benchmark units. Given a multi-session interaction trajectory and a designated anchor turn, a model must perform one of three task families: *judgment*, which asks for the current affective or relational interpretation; *retrieval*, which asks for the historically relevant evidence turns; and *explanation*, which asks the model to justify the interpretation using the relevant history. This design evaluates not only whether a model reaches a plausible answer, but whether it reaches the answer for the right historical reasons.

A-MBER is constructed through a planning-first semi-automated pipeline designed to create long-horizon affective dependencies. Each trajectory begins with persona and role specification, long-horizon event planning, and an emotion arc that unfolds across sessions. Dialogue generation is followed by turn-grounded annotation, anchor definition, candidate item generation, automatic filtering, and human quality verification. The released benchmark contains 34 longitudinal interaction trajectories, 168 sessions, 5,126 dialogue turns, and 2,047 retained benchmark items. Each item is annotated with task family, content type, memory dependency level, reasoning structure, and robustness condition, enabling analysis beyond aggregate accuracy. A stratified human audit of 410 items shows that 93.4% are retained without revision, and the two annotators achieve substantial agreement ($\kappa = 0.82$).

We evaluate ten large language models under three context-access conditions: local-context, full-history, and retrieved-history. This design separates generic model strength from access to historical information. Across models, performance improves substantially when moving from local context to full history, and improves further when models are given retrieved historically relevant turns. The largest gains appear on the subsets A-MBER is designed to stress: long-term implicit affect, high memory-dependency items, multi-hop reasoning, and trajectory-based reasoning. These results suggest that A-MBER measures a capability that is not reducible to local emotion recognition or generic long-context question answering.

Our contributions are threefold. First, we formulate affective memory for emotion recognition as a benchmark target: interpreting the current affective state through remembered multi-session interaction history. Second, we construct A-MBER, a benchmark with anchor-turn-centered tasks, historical evidence grounding, stratified metadata, and human quality verification. Third, we provide a multi-model evaluation under controlled context conditions, showing where historical access and retrieved evidence improve affective interpretation and where current models remain limited.

2 Related Work

Conversational Memory and Long Context. Memory-augmented agents store, retrieve, and update interaction history (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024; Xu et al., 2026; Tan

et al., 2025). Long-term dialogue research studies persistent persona information (Xu et al., 2022a;b), while LoCoMo, LongMemEval, and LoCoMo-Plus evaluate cross-session memory (Maharana et al., 2024; Wu et al., 2024; Li et al., 2026). LongBench and Lost in the Middle distinguish input capacity from effective information use (Bai et al., 2024; Liu et al., 2024). A-MBER focuses on the affective meaning of the present turn under controlled historical access.

Emotion, Causes, and Empathy. Emotion benchmarks support utterance-level and multimodal affect recognition (Li et al., 2017; Hsu et al., 2018; Busso et al., 2008; Zadeh et al., 2018; Poria et al., 2019; Lian et al., 2026). Contextual models incorporate speaker relations, commonsense, and reasoning (Ghosal et al., 2019; 2020; Hu et al., 2021a; Fu et al., 2025). Emotion-cause extraction makes explanatory evidence explicit (Xia & Ding, 2019; Poria et al., 2021; Wang et al., 2024b); empathy and emotional-support benchmarks study understanding and response generation (Rashkin et al., 2019; Liu et al., 2021). A-MBER tests current interpretation, evidence localization, and explanation across a longitudinal personal history.

Grounded Evaluation. Retrieval evaluation separates relevant evidence from faithful answer generation (Saad-Falcon et al., 2024; Niu et al., 2024). Recent studies on multimodal representation learning also explore proxy-based mechanisms for capturing latent structures and personalized grouping patterns (Yao et al., 2024c;b), while efficient sampling methods investigate compact parameterization for adaptive learning (Yao et al., 2024a). LLM judges offer scalable scoring but require human calibration (Liu et al., 2023b; Zheng et al., 2023). A-MBER combines task-specific scores, human verification, and diagnostic subsets. Appendix M reviews these connections, including multimodal robustness, social commonsense, and synthetic-data construction.

3 The A-MBER Benchmark

To systematically evaluate the affective memory capabilities of large models in long-horizon interaction, we construct the A-MBER dataset. Its goal is not to assess whether a model can recognize explicit emotions in a single utterance. Instead, it evaluates whether a model can use the multi-session interaction history of the same user to produce a grounded interpretation of the current affective state or relational state in the current anchor turn.

3.1 Task Formulation and Benchmark Unit

The basic evaluation unit in A-MBER is an anchor-turn-centered benchmark unit. Each unit contains a multi-session interaction trajectory from the same user, an anchor turn to be interpreted, a task instruction, a gold answer, gold evidence, and a set of analysis metadata. Given this information, the model is required to complete one type of task among judgment, retrieval, and explanation tasks.

Formally, let $\mathcal{H}_i = (u_1, \ldots, u_{t_i-1})$ be the ordered pre-anchor interaction history, $a_i = u_{t_i}$ the anchor, and q_i the task instruction. We write an item as

$$d_i = (\mathcal{H}_i, a_i, q_i, y_i, E_i, z_i), \qquad \mathcal{D} = \{d_i\}_{i=1}^{N}, \quad N = 2{,}047, \tag{1}$$

where y_i is the gold answer, E_i the gold evidence turns, and z_i the analysis metadata. For context policy c, a model produces

$$\widehat{y}_i^{(c)} = f_\theta(q_i, a_i, C_c(\mathcal{H}_i, a_i)), \qquad c \in \{\text{local}, \text{full}, \text{retr}\}. \tag{2}$$

Gold answers and evidence are scoring targets; they are not part of this model input.

Judgment tasks require the model to determine the user's current affective state, relational state, or affective orientation in the anchor turn. Retrieval tasks require the model to identify key evidence turns from the historical dialogue that support the current interpretation. Explanation tasks require the model to connect the current state with prior interaction events, recurring problems, support attempts, repaired misunderstandings, or relational changes, and to form an evidence-grounded explanation. The evaluation examines whether the model provides the correct answer for the right historical reasons.

To support stratified analysis, each benchmark item includes corresponding metadata annotations, including content type, memory level, reasoning structure, and robustness tag. Content type distinguishes task categories such as long-range implicit affect, explicit affect, instant emotion, long-range

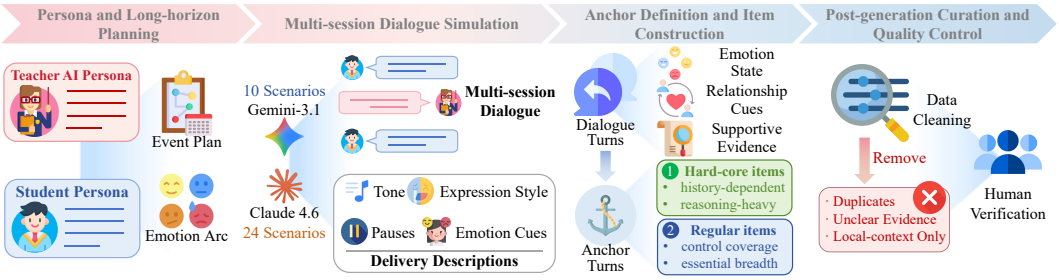

Figure 2: Pipeline of A-MBER.

factual memory, and near-term factual memory. Memory level indicates the degree to which the item depends on historical information. Reasoning structure describes the form of reasoning required by the item, such as direct reasoning, single-hop reasoning, multi-hop reasoning, trajectory-based reasoning, or conflict-resolution reasoning. Robustness tags mark stress-test conditions such as modality-missing, modality-ambiguous, insufficient-evidence, and pseudo-relevant-history conditions.

3.2 Data Construction Pipeline

Because real-world long-horizon psychological support or educational dialogues are difficult to obtain at scale and involve privacy risks, A-MBER adopts a planning-first semi-automated construction pipeline to generate multi-session interaction trajectories. The pipeline shifts the generation target from factual memory and temporal reasoning to present affective interpretation supported by historical evidence. The overall construction pipeline consists of four stages. (see Figure 2)

Persona and Long-horizon Planning. For each scenario, we define the student user, the teacher- or counselor-like AI role, and an event plan and emotion arc that unfold across sessions, so that long-term dependencies are specified before dialogue generation.

Multi-session Dialogue Simulation. Under the constraints of long-horizon planning, Gemini-3.1-pro-preview and Claude 4.6 generate multi-turn dialogue texts. The data also retain structured delivery descriptions, such as tone, pauses, and expressive style, which provide cues relevant to affective interpretation.

Anchor Definition and Candidate Generation. Turn-grounded annotation binds emotional states, relational cues, and supporting historical evidence to observable dialogue turns. We then identify anchor turns and generate candidate questions, gold answers, key explanation points, and gold evidence, covering both high-dependency hard-core items and regular control items.

Post-generation Curation and Quality Control. Candidate items undergo structured field validation, turn ID consistency checks, evidence-answer consistency checks, semantic deduplication, difficulty filtering, and structural pruning. We remove duplicates, unclear evidence references, and samples answerable from local context alone. All 34 trajectories are manually reviewed for scenario coherence, emotion arcs, and cross-session relational development. We further sample 410 items for stratified human verification: 93.4% are retained without modification, 5.1% require minor revision, and 1.5% are removed or rewritten, with substantial annotator agreement (Cohen's κ = 0.82). Details are provided in Appendix A.

3.3 Dataset Statistics and Quality Analysis

The final A-MBER benchmark contains 34 longitudinal interaction trajectories, comprising 168 sessions and 5,126 dialogue turns. On average, each trajectory contains approximately 4.9 sessions and 150.8 dialogue turns. From these trajectories, we construct 2,047 retained benchmark items, with approximately 60.2 items per trajectory and 12.2 items per session. These items are annotated with content type, task family, memory level, reasoning structure, and robustness condition, enabling stratified evaluation beyond aggregate performance. The detailed dataset distribution is provided in Appendix B.

3.4 Analysis Dimensions

A-MBER includes analysis metadata so that results can be interpreted beyond aggregate scores. **Content type** separates long-term implicit emotion, long-term explicit or semi-explicit emotion, instant emotion, long-term factual memory, and near-term factual memory. Long-term implicit emotion items require the model to infer affective meaning from prior interaction patterns rather than from the local wording alone, while instant and near-term items serve as local-control subsets.

Memory level describes the amount of historical dependence required by an item. Level 0 items are largely answerable from the local context; Level 1 items require nearby or same-session evidence; Level 2 items require cross-session evidence; and Level 3 items require integrating multiple historical events, affective shifts, or relationship trajectories.

Reasoning structure describes the form of inference needed to answer an item. Direct and single-hop items depend on one salient cue or evidence turn. Multi-hop items require combining multiple pieces of history. Trajectory-based items require reconstructing how the user's affective state changes over time. Conflict-resolution items require selecting the best-supported interpretation among plausible alternatives. **Robustness tags** mark stress-test cases, including missing or ambiguous modality cues, insufficient evidence, pseudo-relevant history, and other adversarial distractors.

4 Experiments

4.1 Evaluated Models

We evaluate ten large language models from multiple model families, including Gemini, GPT, Claude, Grok, Kimi, GLM, and Qwen. The evaluated model list, version identifiers, context windows, and output limits are provided in Appendix E. All models are evaluated with deterministic decoding. Unless a provider imposes a smaller model-side maximum, we set temperature to 0, top-p to 1.0, and the maximum output length to 64K tokens. For models whose supported maximum output length is below 64K, we use the largest supported value.

We use a diverse set of frontier and widely used API models to test whether A-MBER remains discriminative across model families and whether access to historical context consistently changes performance on affective-memory tasks.

4.2 Context Conditions

A-MBER evaluates each model under three context-access conditions. These conditions differ only in the form of historical information made available to the model. The anchor turn, task instruction, gold answer, gold evidence, and scoring metric remain fixed across conditions.

Local-context. The model receives only the anchor turn or the current session up to the anchor turn. It cannot access earlier sessions. This setting tests how much present affective interpretation can be recovered from local conversational evidence alone.

Full-history context. The model receives the full interaction history before the anchor turn. If the complete history exceeds the model input limit, we apply a fixed truncation rule: the current session up to the anchor turn is always preserved, and earlier historical turns are included in reverse chronological order until the context budget is reached. The final prompt restores all retained turns to their original chronological order. This condition tests whether broad access to raw history is sufficient for long-horizon affective interpretation.

Retrieved-history context. The model receives a retrieved subset of relevant historical turns. We split the interaction history before the anchor turn into turn-level chunks, where each dialogue turn is treated as an independent chunk while preserving its session ID, turn ID, speaker role, and temporal order. For each benchmark item, we construct a retrieval query from the task question, anchor turn, and local context, and use BAAI/bge-m3 as the text embedding model. We retrieve top-$k = 10$ chunks from the historical index. The retrieved chunks are deduplicated, reordered according to their original temporal order, and then concatenated with the task instruction and anchor context. This

condition tests whether models benefit from a compact evidence-focused view of the interaction history rather than the entire raw history.

With query $Q_i = [q_i; a_i; L_i]$, local context L_i, and embedding map e (BAAI/bge-m3), the retrieved view can be written as

$$R_i = \mathrm{TopK}_{u \in \mathcal{H}_i}\left(\frac{e(Q_i)^\top e(u)}{\|e(Q_i)\|\,\|e(u)\|}, 10\right), \qquad C_{\mathrm{retr}} = \mathrm{Chrono}(\mathrm{Dedup}(R_i)) \oplus L_i. \tag{3}$$

Here $\oplus$ denotes context concatenation; turn identifiers retain the link between retrieved text and evidence annotations.

4.3 Evaluation Metrics

Because A-MBER contains multiple task families, scoring is task-specific. Judgment items are scored by exact match against the gold label or option after normalizing option letters and option text. Retrieval items are scored by turn-level overlap between the predicted evidence turns P and gold evidence turns G. We compute precision $= |P \cap G|/|P|$, recall $= |P \cap G|/|G|$, and their harmonic mean:

$$\mathrm{Prec} = \frac{|P \cap G|}{|P|}, \qquad \mathrm{Rec} = \frac{|P \cap G|}{|G|}, \qquad \mathrm{F1} = \frac{2\,\mathrm{Prec}\,\mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}. \tag{4}$$

If the model predicts no evidence turns, precision is defined as 0 unless the gold set is also empty. The same scoring rule is applied across all context conditions for a given item.

For open-ended explanation items, we use a rubric-guided LLM judge because exact string matching cannot capture valid paraphrases of historically grounded explanations. The judge receives the task question, model answer, gold answer, gold rationale, key explanation points, and gold evidence, but not the model identity or context condition. It assigns five criterion scores: core answer correctness (s_c), key point coverage (s_k), rationale alignment (s_r), unsupported claims (s_u), and insufficiency handling (s_i). The final explanation score is:

$$s_{\mathrm{exp}} = .35s_c + .30s_k + .20s_r + .10s_u + .05s_i. \tag{5}$$

The rubric explicitly penalizes generic local interpretations, unsupported historical claims, and answers that miss the specific cross-session mechanism supporting the gold interpretation. The judge returns a JSON object containing all criterion scores, a final score in $[0, 1]$, a verdict, and a short comment. The full judge prompt is provided in Appendix G.

To reduce dependence on an unvalidated automatic evaluator, we conduct a stratified human check of explanation scoring on a held-out sample of model outputs. Human annotators score the same outputs using the rubric dimensions given to the LLM judge, without access to model identity or context condition. We use this sample to check whether the LLM judge preserves the same relative ordering of answers and whether it over-rewards generic but locally plausible explanations. The validation protocol and agreement statistics are reported in Appendix F.

We report both aggregate and stratified results. The primary overall score is computed as the item-level average over all evaluated benchmark items, using each item's task-specific scoring rule. This preserves the released benchmark composition reported in Appendix B, rather than imposing an additional manually chosen weighting over task families. In addition, we report breakdowns by task family, content type, memory level, reasoning structure, and robustness condition. These stratified analyses are central to A-MBER, because the benchmark is designed to reveal where historical affective memory is most useful rather than only whether a model obtains a higher average score.

Let $s_i^{(c)} \in [0, 1]$ denote the task-specific item score. The overall and subset scores, and the paired context gain, are

$$S^{(c)} = \frac{1}{N}\sum_{i=1}^{N} s_i^{(c)}, \quad S_A^{(c)} = \frac{1}{|A|}\sum_{i \in A} s_i^{(c)}, \quad \Delta_A^{c \to c'} = \frac{1}{|A|}\sum_{i \in A}\left(s_i^{(c')} - s_i^{(c)}\right). \tag{6}$$

The subset A is selected by the released metadata rather than by model performance. This keeps the same items paired when comparing context policies.

Table 2: Main results across evaluated models and context conditions. Scores are item-level average overall scores.

Model	Local	Full	Retrieved
Gemini-2.5-F	0.277	0.391	0.453
Gemini-3-Pro	0.301	0.480	0.545
GPT-5.2	0.328	0.505	0.560
Claude-Sonnet	0.349	0.532	0.584
Claude-Opus	**0.371**	**0.565**	**0.613**
GPT-5.4-H	0.352	0.548	0.597
Grok-4.3	0.332	0.501	0.550
Kimi-K2.6	0.322	0.467	0.525
GLM-5.1	0.310	0.451	0.496
Qwen3.6+	0.304	0.443	0.493
Avg.	0.325	0.488	**0.542**

4.4 Reporting Protocol

We log predictions and task-specific scores at the item level, retaining model, context, and metadata identifiers. All aggregate and stratified results use the same scoring script. Context differences are tested by paired bootstrap resampling over benchmark items. The full logging and comparison protocol is provided in Appendix L.

5 Results

5.1 Historical Access Is Necessary but Not Sufficient

The first result is that A-MBER cannot be solved by stronger backbone models using local context alone. Table 2 reports the primary results across ten evaluated models and three context conditions. All models perform substantially worse in the local-context setting than in the history-aware settings, showing that many A-MBER items require information outside the anchor turn or current session. Full-history access improves every model, but retrieved-history access usually improves performance further. This pattern suggests that A-MBER measures not only whether a model can process more context, but whether it can use the historically relevant part of that context.

On average, performance increases from 0.325 under local context to 0.488 under full history and 0.542 under retrieved history. The smaller but consistent retrieved-history gain indicates that evidence selection remains useful even for strong long-context models, while the best retrieved-history score of 0.613 leaves substantial headroom.

This pattern supports the discriminative validity of A-MBER: models differ in their ability to use history, context policies differ in how much usable evidence they expose, and item subsets differ in historical dependence. A benchmark that only measured generic model strength would be less likely to produce these structured gaps across context policy and item metadata. The full 10-model by 12-dimension by 3-context results are provided in Appendix C.

5.2 A-MBER Is Most Diagnostic on History-dependent Affect

The second result is that the largest gains occur on the subsets that A-MBER is designed to stress. Figure 3 reports average results across the ten models by task family, content type, and memory dependency level. Local-context performance is relatively high on instant emotion, near-term factual memory, and Level 0 items, where the anchor turn or nearby context often contains sufficient evidence. In contrast, long-term implicit emotion and high memory-dependency levels show much larger improvements when historical context is available.

This pattern validates the intended structure of the benchmark: if A-MBER were only measuring general emotion recognition, the strongest gains would not concentrate on implicit affect and high-memory subsets. Long-term implicit emotion also remains harder than instant emotion even after retrieval, suggesting that evidence access helps but does not remove the need for historical affective

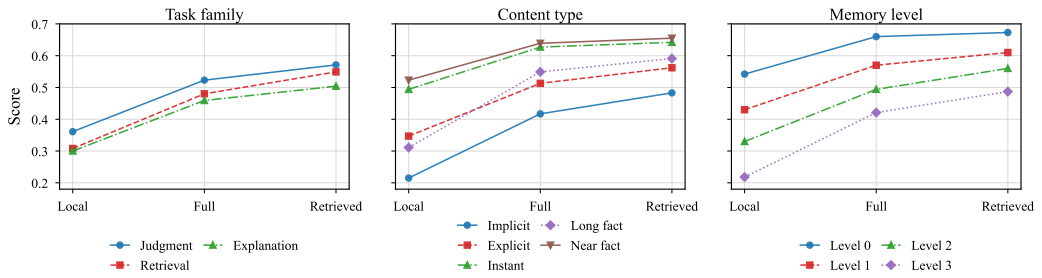

Figure 3: Average stratified results across ten models. Each line shows how a subset changes from local context to full-history and retrieved-history context.

reasoning. The metadata therefore captures meaningful differences in historical dependence rather than merely serving as descriptive annotation.

The task families expose complementary failure modes: judgment tests historically supported interpretation, retrieval tests evidence localization, and explanation requires both the correct interpretation and the correct historical mechanism.

This separation is useful diagnostically because a model may choose the correct affective label while failing to identify the interactional evidence that makes the label warranted. Conversely, a model may retrieve relevant prior turns but still fail to synthesize them into a coherent account of the user's present stance. A-MBER therefore treats answer accuracy, evidence grounding, and explanatory integration as related but distinct aspects of affective memory.

5.3 A-MBER Tests Historical Integration, Not Only Retrieval

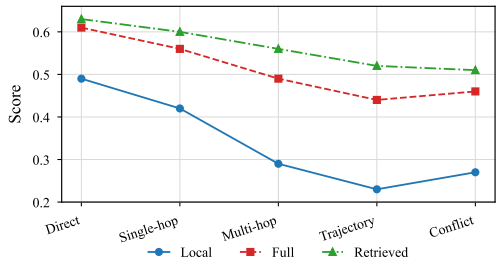

Figure 4: Average results by reasoning structure.

The third result is that historical access helps most when models must integrate evidence across time. Figure 4 breaks down performance by reasoning structure. Direct and single-hop items are easier under local context and show smaller gains from additional history. Multi-hop and trajectory-based items show the largest improvements, indicating that A-MBER tests more than retrieval of isolated facts: it also tests whether models can combine multiple pieces of history into a coherent affective interpretation.

Trajectory-based items make this distinction clearest because they require treating the current state as the endpoint of an interactional path rather than as an isolated label prediction.

The observed gains on these items suggest that history-aware evaluation should not be limited to asking whether a fact from a prior session can be recalled. For affective interaction, the harder problem is often reconstructing how earlier disappointment, repair, hesitation, or trust changes the interpretation of a later utterance.

Paired-bootstrap analysis supports the same conclusion at the item level. The aggregate local-to-full difference yields a 95% confidence interval of [0.148, 0.178], the full-to-retrieved difference yields [0.041, 0.067], and the local-to-retrieved interval is [0.199, 0.235], all excluding zero.

5.4 Robustness and Generator-Source Analyses

The final analysis shows that memory is not uniformly beneficial when evidence is degraded or misleading. Table 3 shows that history improves standard and modality-degraded items, but insufficient-evidence and pseudo-relevant-history subsets remain difficult. This is a desirable stress-test behavior.

Table 3: Average results by robustness condition.

Condition	Local	Full	Retrieved
Standard	0.39	0.58	0.62
Modality-missing	0.31	0.50	0.56
Modality-ambiguous	0.29	0.47	0.53
Insufficient-evidence	0.22	0.36	0.41
Pseudo-relevant-history	0.26	0.42	0.47
Other adversarial	0.30	0.45	0.50

Table 4: Generator-source sensitivity analysis averaged across evaluated models.

Generator source	Local	Full	Retrieved
Gemini-generated subset	0.33	0.51	0.57
Claude-generated subset	0.34	0.52	0.58

A model should not convert every remembered event into a confident explanation; it must also recognize when the available history is weak, ambiguous, or only superficially relevant.

This behavior is especially important for affective memory evaluation. In long-term interactions, many earlier events are emotionally salient but not necessarily relevant to the current anchor turn. A strong model should therefore use history selectively: it should recover prior evidence when it explains the present state, but avoid turning loosely related memories into unsupported causal narratives. The robustness subsets are designed to expose this calibration problem, complementing the main history-access comparison.

We also examine generator-source sensitivity because the trajectories are generated by Gemini- and Claude-family models. The same local-to-full-to-retrieved trend appears for both generator subsets and remains after excluding Gemini- and Claude-family evaluators from the model average, reducing the concern that the main trend is only an artifact of generator-family overlap.

Together, these analyses suggest that the benchmark is not driven by a single artifact, such as one generator family, one context policy, or one easy subset. The main trend persists across model families and data sources, while the stratified results reveal where historical affective reasoning remains difficult.

6 Conclusion

A-MBER evaluates whether models can interpret a current affective or relational state using the relevant multi-session history. Its judgment, retrieval, and explanation tasks separate answer correctness, evidence localization, and historical integration. Across ten models, full history improves on local context, and retrieved history provides further gains, especially for implicit affect, high-memory items, multi-hop reasoning, and affective trajectories. Insufficient evidence and pseudo-relevant history remain challenging, highlighting the need for calibrated use of remembered events.

These findings motivate memory evaluation beyond factual recall and context length. The benchmark is based on generated trajectories and rubric-guided explanation scoring; naturalistic interactions, richer multimodal evidence, and explicit memory-writing and updating systems remain directions for future evaluation. Detailed limitations appear in Appendix H.

Reproducibility Statement

The appendices provide the human verification protocol, dataset composition, full model-by-dimension results, model and retrieval settings, judge validation, and the complete explanation-scoring prompt. Additional formal definitions clarify context construction, score aggregation, and paired comparisons.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024.

Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 370–379, 2019.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6168–6173, 2019.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14388–14411, 2024.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3119–3137, 2024.

Emily M Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the association for computational linguistics*, 6:587–604, 2018.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew Gormley. Unlimiformer: Long-range transformers with unlimited length input. volume 36, pp. 35522–35543, 2023.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4762–4779, 2019.

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.

Mao Yan Chen, Siheng Li, and Yujiu Yang. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1063–1074, 2022.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 646–651, 2018.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3014–3026, 2022.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 4079–4095, 2023.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4443–4458, 2020.

Zixiang Ding, Rui Xia, and Jianfei Yu. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3161–3170, 2020.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

Huawen Feng, Junlong Liu, Junhao Zheng, Haibin Chen, Xichen Shang, and Qianli Ma. Joint constrained learning with boundary-adjusting for emotion-cause pair extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1118–1131, 2023.

Sarah E Finch and Jinho D Choi. Convosense: Overcoming monotonous commonsense inferences for conversational ai. *Transactions of the Association for Computational Linguistics*, 12:467–483, 2024.

Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. There are a thousand hamlets in a thousand people's eyes: Enhancing knowledge-grounded dialogue with personal memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3901–3913, 2022.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6748–6761, 2025.

Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2843–2853, 2022.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. Peacok: Persona commonsense knowledge for consistent and engaging narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6569–6591, 2023.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 2470–2481, 2020.

Omer Goldman, Alon Jacovi, Aviv Slobodkin, Aviya Maimon, Ido Dagan, and Reut Tsarfaty. Is it really long context if all you need is retrieval? towards genuinely difficult long context nlp. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16576–16586, 2024.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1726–1736, 2024.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2594–2604, 2018a.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2122–2132, 2018b.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7042–7052, 2021a.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10835–10852, 2023.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7837–7851, 2022.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021b.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 7360–7370, 2020.

DongJin Jeong and Jinyeong Bak. Conversational emotion-cause pair extraction with guided mixture of experts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3288–3298, 2023.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 7036–7050, 2024.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 397–406, 2019.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4148–4164, 2022.

Dongjin Kang, Sunghwan Mac Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 15232–15261, 2024.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4110–4124, 2021.

Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. Commonsense-augmented memory construction and management in long-term conversations via context-aware persona refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 104–123, 2024.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2227–2240, 2021.

Wongyu Kim, Youbin Ahn, Donghyun Kim, and Kyong-Ho Lee. Emp-rft: Empathetic response generation via recognizing feature transitions between utterances. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4118–4128, 2022.

Philippe Laban, Alexander Richard Fabbri, Caiming Xiong, and Chien-Sheng Wu. Summary of a haystack: A challenge to long-context llms and rag systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9885–9903, 2024.

Andrew Lee, Jonathan K Kummerfeld, Larry An, and Rada Mihalcea. Empathy identification systems are not accurately accounting for context. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1686–1695, 2023.

Joosung Lee and Wooin Lee. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 5669–5679, 2022.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1976–1988, 2024.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, 2017.

Yifei Li, Weidong Guo, Lingling Zhang, Rongman Xu, Muye Huang, Hui Liu, Lijiao Xu, Yu Xu, and Jun Liu. Locomo-plus: Beyond-factual cognitive memory evaluation framework for llm agents. pp. 25085–25100, 2026.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 121–132, 2019.

Junlong Liu, Xichen Shang, and Qianli Ma. Pair-based joint encoding with relational graph convolutional networks for emotion-cause pair extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5339–5351, 2022.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8404–8419, 2023a.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3469–3483, 2021.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. volume 36, pp. 46534–46594, 2023.

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, 2024.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 8968–8979, 2020.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. Pal to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12254–12271, 2023.

Saif M Mohammad. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278, 2022.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. Glucose: Generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4569–4586, 2020.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Seong-Il Park and Jay-Yoon Lee. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 12:1686–1702, 2024.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7682–7696, 2024.

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, 2019.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5811–5826, 2020.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. Timedial: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 7066–7076, 2021.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 5835–5847, 2021.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2289–2299, 2018a.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 463–473, 2018b.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5370–5381, 2019.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7881–7892, 2020.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276, 2020.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1551–1560, 2021.

Tao Shi and Shao-Lun Huang. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 14752–14766, 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 8371–8384, 2024.

Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. Language models (mostly) do not consider emotion triggers when predicting emotion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 603–614, 2024.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 658–670, 2023.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4455–4480, 2024.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 308–319, 2022.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3712–3724, 2024a.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th international workshop on semantic evaluation (SemEval-2024)*, pp. 2039–2050, 2024b.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, et al. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5627–5646, 2024c.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. volume 36, pp. 74530–74543, 2023a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023b.

Zheng Wang, Zhongyang Li, Zeren Jiang, Dandan Tu, and Wei Shi. Crafting personalized agents through retrieval-augmented generation on editable memory graphs. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 4891–4906, 2024d.

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. volume 35, pp. 24824–24837, 2022.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. Personalized response generation via generative split memory network. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 1956–1970, 2021.

Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1003–1012, 2019.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4643–4663, 2024.

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, 2022a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, 2022b.

Jiamin Yang and David Jurgens. Modeling empathetic alignment in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3127–3148, 2024.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. An iterative associative memory model for empathetic response generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3081–3092, 2024.

Jiawei Yao, Chuming Li, and Canran Xiao. Swift sampler: Efficient learning of sampler by 10 parameters. *Advances in Neural Information Processing Systems*, 37:59030–59053, 2024a.

Jiawei Yao, Qi Qian, and Juhua Hu. Customized multiple clustering via multi-modal subspace proxy learning. *Advances in Neural Information Processing Systems*, 37:82705–82725, 2024b.

Jiawei Yao, Qi Qian, and Juhua Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14066–14075, 2024c.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 82–95, 2024.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2236–2246, 2018.

Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pp. 44–52, 2018.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7395–7408, 2023.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2386–2398, 2024a.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 13395–13412, 2024b.

Xiaoheng Zhang and Yang Li. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13099–13110, 2023.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, et al. Esc-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 15785–15810, 2024a.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 2608–2618, 2021.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. volume 36, pp. 46595–46623, 2023.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11325–11345, 2024.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5808–5820, 2022.

Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 165–176, 2019.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6556–6566, 2020.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1714–1729, 2023a.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8223–8237, 2023b.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

A HUMAN QUALITY VERIFICATION PROTOCOL

To examine the reliability of the generated benchmark items, we conducted a two-stage human quality verification procedure after automatic post-generation curation, assessing whether the retained items were answerable from the dialogue history, whether the gold answers and gold evidence were well supported, and whether the analysis metadata accurately reflected the intended evaluation dimensions.

In the first stage, annotators reviewed all 34 longitudinal interaction trajectories included in the benchmark. This review focused on the overall quality and coherence of the generated interaction histories rather than on individual benchmark questions alone. For each trajectory, annotators examined whether the persona setting remained stable across sessions, whether major events developed in a coherent order, whether the user's affective trajectory was plausible, and whether cross-session changes in trust, hesitation, conflict, repair, or emotional stance were understandable from the dialogue.

The trajectory-level review was used to identify global construction problems that could affect multiple downstream benchmark items. Examples included inconsistent persona information, abrupt emotional transitions without sufficient interactional support, unclear event references across sessions, or implausible shifts in the user–agent relationship. Trajectories with such issues were flagged for revision before item-level auditing.

In the second stage, we conducted stratified item-level verification on a sampled subset of the retained benchmark items. Since the final benchmark contains 2,047 items, a full double-annotated review of every item would be costly. We therefore sampled 410 items, corresponding to 20.0% of the benchmark.

The sample was stratified by the main analysis dimensions used in A-MBER, including task family, content type, memory dependency level, reasoning structure, and robustness condition. This ensured that the audited subset covered not only common item types, but also diagnostically important cases such as high-memory-dependency items, trajectory-based reasoning items, insufficient-evidence items, pseudo-relevant-history items, and modality-ambiguous items.

For low-frequency but theoretically important stress-test categories, we used moderate oversampling to ensure that these conditions were represented in the audit. Sampling weights were recorded so that the audit results could be interpreted with respect to both the sampled subset and the overall benchmark distribution.

Each sampled item was evaluated along four dimensions.

Answerability. Annotators judged whether the question could be answered from the provided dialogue history and anchor turn. Items were marked as problematic if the correct answer required information not present in the observable interaction history.

Gold-answer consistency. Annotators examined whether the gold answer was consistent with the anchor turn, the relevant historical context, and the intended task type. For judgment items, this involved checking whether the selected label or option matched the most defensible interpretation. For explanation items, this involved checking whether the gold answer captured the core affective or relational interpretation.

Evidence sufficiency. Annotators assessed whether the gold evidence was sufficient to support the gold answer. Evidence was considered sufficient when the cited turns directly or cumulatively supported the target interpretation. Items were flagged if the evidence was only superficially related, too broad, too narrow, or insufficient to distinguish the gold answer from plausible alternatives.

Metadata validity. Annotators checked whether the item metadata matched the actual reasoning demand of the item. This included content type, memory dependency level, reasoning structure, and robustness or adversarial tags. Items were flagged when, for example, a supposedly high-memory-dependency item could be answered from local context alone, or when a trajectory-based item only required a single isolated fact.

Two annotators independently reviewed each sampled item. For every item, annotators assigned one of three overall decisions: retain, minor revision, or remove/rewrite.

An item was marked as retain if the question was answerable, the gold answer was supported, the gold evidence was appropriate, and the metadata was broadly valid.

An item was marked as minor revision if the core item was usable but required limited correction, such as narrowing the evidence range, revising an overly general gold answer, clarifying the wording of the question, or adjusting metadata labels.

An item was marked as remove/rewrite if it had a fundamental issue, such as insufficient historical evidence, an unsupported gold answer, unclear anchor-turn interpretation, severe metadata mismatch, or answerability from local context despite being designed as a long-memory item.

After independent annotation, disagreements were resolved through discussion. When the two annotators could not reach agreement, the item was reviewed by an additional adjudicator. All items marked as minor revision or remove/rewrite were revised, removed, or regenerated before inclusion in the final benchmark.

Based on the human audit, 383 out of 410 sampled items were accepted without revision, corresponding to 93.4% of the audited subset. Twenty-one items, or 5.1%, required minor revision. The most common revision cases involved evidence spans that were broader than necessary, gold answers that required more specific wording, or metadata labels that needed adjustment. Six items, or 1.5%, were removed or rewritten because they lacked sufficient historical evidence or could be answered from local context alone.

Inter-annotator agreement was calculated on the three-way retain/revise/remove decision. The two annotators achieved a Cohen's κ of 0.82, indicating substantial agreement. This result suggests that the verification criteria were sufficiently well defined for consistent human judgment. Detailed human quality verification results are summarized in Table 5.

Table 5: Human quality verification results.

Decision	Count	Percentage
Retain without revision	383	93.4%
Minor revision	21	5.1%
Remove or rewrite	6	1.5%
Total audited items	410	100.0%

B Dataset Statistics

The released A-MBER benchmark contains 34 longitudinal interaction trajectories, comprising 168 sessions and 5,126 dialogue turns in total. From these trajectories, we derive 2,047 retained benchmark items. Table 6 reports the detailed distribution of the released benchmark across content type, task family, robustness condition, memory level, and reasoning structure. These metadata support the stratified analyses reported in the main experiments.

Table 6: Detailed distribution of A-MBER benchmark items. Counts are reported over the 2,047 retained benchmark items. Content type, memory level, and reasoning structure each define a full partition of the benchmark items. Task family and robustness condition describe the released evaluation composition used in the benchmark.

Dimension	Category	Count	Percentage
Content type	Long-term implicit emotion	1,019	49.8%
	Long-term explicit / semi-explicit emotion	337	16.5%
	Instant emotion	157	7.7%
	Long-term fact	286	14.0%
	Near-term fact	248	12.1%
Task family	Judgment	982	48.0%
	Retrieval	371	18.1%
	Explanation	529	25.8%
Robustness condition	Modality-missing / ambiguous	165	8.1%
Memory level	Level 0	223	10.9%
	Level 1	104	5.1%
	Level 2	512	25.0%
	Level 3	1,208	59.0%
Reasoning structure	Direct	221	10.8%
	Single-hop	634	31.0%
	Multi-hop	671	32.8%
	Conflict-resolution	247	12.1%
	Trajectory-based	274	13.4%

C Detailed Model-by-Dimension Results

Table 7 reports the full model-by-dimension results underlying the aggregate and stratified analyses in the main text. The twelve diagnostic dimensions correspond to the three task families, five content types, and four memory-dependency levels used throughout A-MBER. Scores are item-level averages under the same scoring rules as Section 4.

Table 7: Detailed $10\times12\times3$ model-by-dimension results. Each panel reports one context condition. Abbreviations: Judg.=judgment, Retr.=retrieval, Expl.=explanation, Impl.=implicit emotion, Explct.=explicit emotion, Inst.=instant emotion, L-Fact=long-term fact, N-Fact=near-term fact, L0–L3=memory levels 0–3.

Local context	Judg.	Retr.	Expl.	Impl.	Explct.	Inst.	L-Fact	N-Fact	L0	L1	L2	L3
Gemini-2.5-F	0.324	0.266	0.267	0.174	0.315	0.458	0.270	0.490	0.502	0.396	0.291	0.183
Gemini-3-Pro	0.337	0.293	0.278	0.202	0.330	0.471	0.297	0.502	0.526	0.410	0.314	0.199
GPT-5.2	0.367	0.307	0.308	0.219	0.345	0.501	0.311	0.529	0.543	0.435	0.332	0.216
Claude-Sonnet	0.377	0.332	0.320	0.230	0.371	0.510	0.333	0.540	0.563	0.449	0.344	0.241
Claude-Opus	0.402	0.345	0.332	0.256	0.380	0.533	0.345	0.561	0.578	0.461	0.370	0.251
GPT-5.4-H	0.384	0.325	0.326	0.234	0.371	0.513	0.334	0.544	0.558	0.455	0.348	0.245
Grok-4.3	0.362	0.318	0.303	0.224	0.351	0.502	0.316	0.524	0.551	0.432	0.341	0.225
Kimi-K2.6	0.363	0.303	0.301	0.211	0.347	0.492	0.304	0.525	0.536	0.433	0.329	0.212
GLM-5.1	0.347	0.299	0.287	0.206	0.335	0.477	0.303	0.508	0.536	0.420	0.314	0.211
Qwen3.6+	0.348	0.290	0.286	0.199	0.326	0.482	0.291	0.512	0.527	0.410	0.318	0.199

Table 8: Detailed model-by-dimension results under the full-history context condition.

Full context	Judg.	Retr.	Expl.	Impl.	Explct.	Inst.	L-Fact	N-Fact	L0	L1	L2	L3
Gemini-2.5-F	0.443	0.408	0.380	0.347	0.439	0.547	0.478	0.561	0.588	0.493	0.421	0.345
Gemini-3-Pro	0.521	0.471	0.458	0.412	0.503	0.626	0.540	0.636	0.652	0.566	0.488	0.410
GPT-5.2	0.533	0.499	0.474	0.426	0.531	0.638	0.565	0.651	0.676	0.583	0.502	0.438
Claude-Sonnet	0.563	0.516	0.489	0.456	0.545	0.665	0.582	0.676	0.694	0.600	0.532	0.452
Claude-Opus	0.585	0.536	0.524	0.475	0.576	0.685	0.611	0.699	0.715	0.634	0.551	0.487
GPT-5.4-H	0.566	0.532	0.503	0.467	0.558	0.676	0.596	0.681	0.711	0.614	0.546	0.469
Grok-4.3	0.538	0.488	0.472	0.426	0.526	0.637	0.554	0.653	0.667	0.586	0.506	0.427
Kimi-K2.6	0.504	0.467	0.441	0.403	0.497	0.606	0.537	0.619	0.649	0.555	0.474	0.409
GLM-5.1	0.497	0.450	0.433	0.388	0.479	0.602	0.517	0.616	0.633	0.537	0.470	0.390
Qwen3.6+	0.486	0.447	0.424	0.377	0.482	0.589	0.520	0.606	0.621	0.540	0.456	0.389

Table 9: Detailed model-by-dimension results under the retrieved-history context condition.

Retrieved context	Judg.	Retr.	Expl.	Impl.	Explct.	Inst.	L-Fact	N-Fact	L0	L1	L2	L3
Gemini-2.5-F	0.506	0.476	0.440	0.415	0.489	0.577	0.519	0.589	0.602	0.543	0.491	0.413
Gemini-3-Pro	0.570	0.557	0.508	0.481	0.569	0.642	0.597	0.656	0.678	0.612	0.558	0.493
GPT-5.2	0.591	0.565	0.514	0.502	0.574	0.659	0.603	0.671	0.687	0.619	0.578	0.498
Claude-Sonnet	0.605	0.578	0.542	0.513	0.598	0.673	0.626	0.688	0.701	0.647	0.590	0.525
Claude-Opus	0.622	0.609	0.557	0.542	0.616	0.700	0.646	0.706	0.732	0.662	0.621	0.544
GPT-5.4-H	0.619	0.590	0.550	0.524	0.607	0.685	0.629	0.702	0.713	0.658	0.605	0.526
Grok-4.3	0.575	0.558	0.509	0.492	0.568	0.643	0.601	0.658	0.685	0.618	0.562	0.498
Kimi-K2.6	0.561	0.534	0.493	0.470	0.544	0.633	0.575	0.647	0.661	0.593	0.552	0.471
GLM-5.1	0.534	0.516	0.468	0.442	0.530	0.603	0.561	0.621	0.633	0.579	0.522	0.454
Qwen3.6+	0.535	0.511	0.461	0.449	0.521	0.609	0.554	0.613	0.640	0.569	0.525	0.447

D Qualitative Case Study

Table 10 shows a real A-MBER item from the released benchmark package. The item is useful as a qualitative example because the anchor turn is locally an ordinary academic question, while the relevant history reveals it as an avoidance strategy. This illustrates why A-MBER evaluates historical affective interpretation rather than only surface emotion recognition.

Table 10: Representative A-MBER case study from item U_Q019. The example is shortened for readability.

Field	Content
Metadata	`conversation_id`: `c_psych_teacher_student_001`; `question_id`: `Q019`; content type: long-term implicit emotion; task family: judgment; memory level: Level 3; reasoning structure: single-hop.
Question	How does the student's sudden shift to APA citation formatting contrast with his approach to academic deadlines in the previous meeting?
Anchor turn	Mason says: "Understood. Actually, while I'm here, I wanted to ask about the essay draft feedback. For the APA citation formatting, do we need to include the DOI for all online journal articles?" The voice annotation marks a sudden shift to a brisker pace with clear articulation on "APA citation formatting."
Relevant history	In an earlier meeting, Mason requested a short extension and described it as "just a minor time-management glitch." Immediately before the anchor turn, the teacher says that coffee is not a solution and that Mason needs actual rest to process the material.
Gold interpretation	Previously, Mason minimized academic difficulty as a minor deadline issue; here, he uses a highly specific academic question as a conversational shield to deflect personal scrutiny about sleep and fatigue.
Local-context failure mode	A local answer can treat the turn as a sincere formatting question or ordinary task focus, missing the pragmatic contrast with earlier minimization.
Retrieved-history interpretation	With the retrieved evidence, the answer can connect the abrupt APA question to the earlier "minor glitch" framing and identify the repeated pattern of shifting from uncomfortable personal scrutiny to safe academic details.

E Model and Retrieval Configuration

Table 11 lists the large language models evaluated in the experiments. All models are evaluated with deterministic decoding. We set temperature to 0 and top-p to 1.0. The evaluation output limit is set to 64K tokens whenever supported by the provider; when a model supports a smaller maximum output length, we use the provider-side maximum.

Table 11: The list of evaluated large language models. "Context / Eval Max Output" reports the provider-reported context window and the output cap used in our evaluation.

Abbreviation	Version ID / Configuration	Developer	Context / Output
	Reasoning-Oriented API Models		
Gemini-3-Pro	`gemini-3-pro-preview`	Google	1M / 64K
GPT-5.4-High	`gpt-5.4` with high reasoning	OpenAI	1M / 64K
GPT-5.2	`gpt-5.2`	OpenAI	400K / 64K
Claude-Opus-4.1	`claude-opus-4-1-20250805`	Anthropic	200K / 32K
Claude-Sonnet-4	`claude-sonnet-4-20250514`	Anthropic	200K / 64K
Grok-4.3	`grok-4.3`	xAI	1M / 64K
GLM-5.1	`glm-5.1`	Z.ai	200K / 64K
Qwen3.6-Plus	`qwen3.6-plus`	Alibaba Cloud	1M / 64K
Kimi-K2.6	`kimi-k2.6`	Moonshot AI	256K / 64K
	Efficient / Fast API Models		
Gemini-2.5-Flash	`gemini-2.5-flash`	Google	1M / 64K

For retrieved-history context, we index the history before the anchor turn at the turn level. Each dialogue turn is treated as an independent chunk and is stored together with its trajectory ID, session ID, turn ID, speaker role, and original temporal position. For each benchmark item, the retrieval query is formed by concatenating the task question, the anchor turn, and the local current-session context. We embed both queries and historical chunks with BAAI/bge-m3 and retrieve the top-$k = 10$ chunks by cosine similarity. Retrieved chunks are deduplicated by turn ID and then reordered according to their original chronological order before being inserted into the model prompt.

For full-history context, if the complete pre-anchor history exceeds the model input budget, we preserve the local current-session context first and then add earlier turns in reverse chronological order until the budget is reached. The retained turns are finally presented to the model in chronological order. This rule ensures that the anchor-proximal context is always available while still allowing as much earlier history as possible to be included.

F LLM Judge Validation

We validate the rubric-guided LLM judge on a stratified sample of 180 open-ended explanation outputs. The sample is drawn across task difficulty, content type, memory level, reasoning structure, context condition, and model family, so that the validation set includes both easy local explanations and high-memory cases requiring cross-session evidence. Human annotators score each sampled output using the same five rubric dimensions used by the LLM judge: core answer correctness, key point coverage, rationale alignment, unsupported claims, and insufficiency handling.

The validation focuses on two questions. First, we test whether the LLM judge preserves the relative quality ordering of answers, especially between generic local explanations and historically grounded explanations. Second, we examine whether the judge over-rewards fluent but unsupported historical claims. Table 12 reports rank correlation between human and LLM final scores, mean absolute difference in final score, and coarse verdict agreement. Disagreements are inspected manually and grouped into cases such as overly generous scoring for vague answers, overly harsh scoring for concise but correct answers, and disagreement over whether a historical dependency is necessary.

Table 12: Human validation of the LLM judge on explanation outputs.

Metric	Value
Sampled explanation outputs	180
Spearman correlation	0.78
Pearson correlation	0.74
Mean absolute difference	0.083
Coarse verdict agreement	81.7%

This validation does not make the LLM judge equivalent to human annotation. Rather, it serves as a calibration check for using the judge as a scalable scoring tool in the full benchmark evaluation. We therefore interpret LLM-judged explanation scores together with exact-match judgment scores, retrieval F1, item-level logs, and the human quality verification of benchmark items.

G LLM JUDGE PROMPT

LLM Judge Prompt for Open-ended A-MBER Evaluation

```
You are a strict evaluator for an emotion-memory benchmark.
Your job is to judge whether a model answer is correct for an open-ended benchmark
    question.

Score the answer with a weighted rubric. The dimensions are not equally important.

Primary dimensions:
1. core_answer_correct (highest importance)
   - Measures whether the answer captures the core meaning of the gold answer or an
     acceptable answer.
   - Use a score between 0.0 and 1.0.
   - Suggested anchors:
     - 1.0 = fully captures the core answer
     - 0.75 = mostly correct, minor omission
     - 0.5 = partially correct or too vague
     - 0.25 = weakly related but misses the key claim
     - 0.0 = incorrect
   - For long-memory or explanation questions, do NOT give a high score to an answer
      that only gives the right broad vibe while missing the specific historical
     mechanism that makes the gold answer correct.
   - If several explanations are superficially plausible, do NOT give a high score
     unless the answer identifies why the gold explanation is better supported than
     the main alternative.

2. key_point_coverage (high importance)
   - Measures how many key_explanation_points are covered.
   - Use a score between 0.0 and 1.0.
   - Allow paraphrases and compressed wording.
   - Suggested anchors:
     - 1.0 = covers nearly all important points
     - 0.75 = covers most points
     - 0.5 = covers about half
     - 0.25 = covers only a small part
     - 0.0 = covers none

Secondary dimensions:
3. rationale_alignment (medium importance)
   - Measures whether the answer follows the reasoning direction of gold_rationale.
   - Use a score between 0.0 and 1.0.
   - Suggested anchors:
     - 1.0 = strongly aligned
     - 0.75 = mostly aligned
```

```
     - 0.5 = broadly aligned but missing a key reasoning step
     - 0.25 = weak alignment
     - 0.0 = inconsistent
   - If the gold answer depends on a particular earlier mismatch, repair,
     reinterpretation, or cross-session causal link, an answer that omits that
     dependency should score low here even if the local interpretation sounds
     plausible.
   - If the gold answer depends on weighing competing explanations, an answer that
     names the gold one but does not rule out or distinguish the strongest rival
     explanation should score only moderate here.

4. unsupported_claims (penalty-sensitive, lower importance than correctness)
   - Measures whether the answer adds unsupported claims, hallucinations, or
     contradictions.
   - Use a score between 0.0 and 1.0.
   - Interpret this as factual discipline:
     - 1.0 = no unsupported claims
     - 0.75 = very minor overreach
     - 0.5 = noticeable but not fatal overreach
     - 0.25 = substantial unsupported content
     - 0.0 = clear hallucination or contradiction

5. insufficiency_handling (conditional dimension)
   - Only matters when the correct response should acknowledge insufficient
     information.
   - Use a score between 0.0 and 1.0.
   - If insufficiency is NOT relevant for this item, set this to 1.0 by default
     rather than penalizing.
   - Suggested anchors:
     - 1.0 = handles insufficiency correctly
     - 0.5 = recognizes uncertainty but still overcommits
     - 0.0 = confidently invents unsupported content when insufficiency should have
     been acknowledged

Weighted scoring guidance:
- final_score must be a weighted combination on a 0.0 to 1.0 scale.
- Use this weighting:
  - core_answer_correct: 0.35
  - key_point_coverage: 0.30
  - rationale_alignment: 0.20
  - unsupported_claims: 0.10
  - insufficiency_handling: 0.05
- Do NOT treat all dimensions as equally important.
- Do NOT collapse the final score to only 0.0 / 0.5 / 1.0.
- Use finer-grained values when justified, for example 0.15, 0.3, 0.45, 0.6, 0.7,
    0.8, 0.9.
- Reserve 1.0 for answers that are genuinely excellent.
- Reserve 0.0 for answers that are genuinely wrong or unusable.

Additional judging rules:
- Respect acceptable paraphrases.
- Do not reward unsupported claims even if part of the answer is correct.
- If the answer is concise but correct, do not penalize brevity.
- If the answer misses a key historical dependency in a long-memory question, reduce
     both core_answer_correct and rationale_alignment appropriately.
- For explanation questions, generic descriptions such as "the student is guarded",
    "the student is stressed", or "the student is minimizing" are not enough for
    high scores unless the answer explains what earlier interaction pattern makes
    that interpretation correct here.
- For explanation or judgment questions that hinge on competing readings, generic
    answers such as "the student is guarded", "the student is stressed", or "the
    student is minimizing" are not enough for high scores unless the answer explains
     why that reading is stronger than the main competing interpretation.
```

```
- For gold-evidence evaluation, hold the answer to a stricter standard: if the
    supplied evidence should allow a more specific answer and the model still stays
    at a shallow local reading, lower key_point_coverage and rationale_alignment
    noticeably.
- If the model clearly selects the correct option semantically but answers with the
    full option text instead of the letter, do not treat that as incorrect in your
    open-ended judgment.
- If the model gives the right letter or option text for a judgment item but cannot
    justify it in the open-ended comment, core_answer_correct may still be high, but
     key_point_coverage and rationale_alignment should stay lower.
- If the answer mainly repeats a late explicit recap, summary line, or surface
    conclusion without reconstructing the earlier interactional path that made it
    true, cap core_answer_correct and rationale_alignment below the top band.

Return JSON only with this schema:
{
  "core_answer_correct": 0.0 to 1.0,
  "key_point_coverage": 0.0 to 1.0,
  "rationale_alignment": 0.0 to 1.0,
  "unsupported_claims": 0.0 to 1.0,
  "insufficiency_handling": 0.0 to 1.0,
  "final_score": 0.0 to 1.0,
  "verdict": "short label",
  "comment": "short explanation"
}
```

H Limitations and Scope

A-MBER has several limitations. First, the benchmark is constructed from semi-automatically generated long-horizon interactions rather than naturally collected human-agent histories. Although we use planning constraints, automatic filtering, and human quality verification, generated trajectories may still differ from real longitudinal interaction in topic distribution, emotional subtlety, and conversational messiness.

Second, the dialogue trajectories are generated by Gemini- and Claude-family models. This introduces a possible generator-source bias, especially when evaluator models come from the same model families. We therefore report generator-source sensitivity and non-overlap analyses, but these checks cannot fully remove all effects of synthetic data generation.

Third, explanation scoring relies on a rubric-guided LLM judge. The judge is designed to penalize generic local interpretations and unsupported historical claims, and we conduct human validation on a stratified sample of explanation outputs. Nevertheless, LLM-as-judge evaluation can still introduce calibration bias, especially for open-ended explanations with multiple defensible phrasings.

Finally, A-MBER evaluates three context policies: local context, full history, and retrieved history. These settings isolate the role of historical access, but they do not exhaust the design space of memory systems. Future work should evaluate explicit memory-writing, memory-updating, forgetting, conflict resolution over persistent memory stores, and end-to-end deployed agents interacting with users over longer periods.

I Formal View of Affective-Memory Evaluation

This section expands the task and context definitions in the main text. The notation expresses the benchmark protocol already described above; it does not introduce an additional model, training objective, or experimental condition.

I.1 History, Anchors, and Evidence

A longitudinal trajectory is an ordered sequence of sessions, and a session is an ordered sequence of dialogue turns. Let $\tau(u)$ denote the temporal position of a turn and $\sigma(u)$ its session identifier. For

anchor a_i, the admissible history is

$$\mathcal{H}_i = \{u : \tau(u) < \tau(a_i)\}, \qquad E_i \subseteq \mathcal{H}_i. \tag{7}$$

The ordering matters: an earlier statement of disappointment, a later repair attempt, and the current cautious response are not interchangeable evidence. The model must interpret the anchor at its own point in the interaction. A statement made after the anchor cannot establish what was recoverable before that statement occurred.

An item couples an answer target with an evidence target. For judgment, the constrained answer space makes label correctness easy to measure, but a correct label alone does not demonstrate that the underlying historical mechanism was understood. For retrieval, the target is a set of turn identifiers. For explanation, the target includes the interpretation and the key connections between historical events and the present state. These output spaces make the three tasks complementary rather than interchangeable.

The metadata vector can be written as

$$z_i = (t_i, c_i, m_i, r_i, b_i), \tag{8}$$

where t_i is task family, c_i content type, m_i memory level, r_i reasoning structure, and b_i robustness condition. Memory level concerns the historical dependence of an item; reasoning structure concerns the operation performed on the evidence. They should therefore be read as separate annotations. In particular, a cross-session or high-memory item can still have a single decisive supporting contrast, as illustrated by the case study in Appendix D.

I.2 What the Context Comparison Holds Fixed

For a fixed item, context policies change the information exposed to the model. The task question, anchor, reference answer, evidence annotation, and scoring rule stay fixed. Let $I_i^{(c)}$ be the model input and T_i the evaluation target:

$$I_i^{(c)} = (q_i, a_i, C_c(\mathcal{H}_i, a_i)), \qquad T_i = (y_i, E_i). \tag{9}$$

This separation is useful because a change in score can then be interpreted as a change under historical access for the same evaluation item. Gold evidence is used to evaluate grounding; the retrieved-history setting uses embedding retrieval rather than supplying the gold evidence as an oracle context.

Under full-history access, the input budget can prevent every earlier turn from being included. Our protocol preserves the current session and fills the remaining budget from recent to older turns, before restoring chronological order. Thus, "full history" names the context policy; it does not imply that arbitrarily long trajectories fit into every model. Comparisons across models should be interpreted together with their reported context budgets.

For retrieved history, similarity selects candidate turns, while chronological reordering preserves their interactional sequence. These steps address different needs: similarity makes evidence selection compact, and ordering makes changes across time interpretable. Neither step guarantees that the selected turns contain all evidence needed for a correct explanation. The remaining error under retrieved context is therefore consistent with failures of evidence selection, evidence integration, or both.

I.3 Separating Access Gains from Selection Gains

The main context comparison admits a simple decomposition. For the same set of evaluated items A,

$$\Delta_A^{\text{access}} = S_A^{(\text{full})} - S_A^{(\text{local})}, \tag{10}$$

$$\Delta_A^{\text{selection}} = S_A^{(\text{retr})} - S_A^{(\text{full})}, \tag{11}$$

$$\Delta_A^{\text{total}} = \Delta_A^{\text{access}} + \Delta_A^{\text{selection}}. \tag{12}$$

These are descriptive score differences, not separate causal estimates of internal model mechanisms. Full and retrieved contexts differ in length, distractors, and which evidence is present. A positive selection gain means that the compact retrieved view is more useful under the reported protocol; it does not by itself establish that a particular memory architecture is superior.

Using the rounded averages in Table 2, the differences are $0.488 - 0.325 = 0.163$, $0.542 - 0.488 = 0.054$, and $0.542 - 0.325 = 0.217$. These quantities restate the main results and introduce no additional observations. The rounded values also make clear why aggregate improvements should be accompanied by subset analysis: the same average can conceal very different behavior on instant emotion, implicit emotion, and high-memory items.

J Reading the Diagnostic Results

J.1 Answer Correctness and Historical Grounding

Affective interpretations are often plausible at more than one level of specificity. A model may correctly describe a student as stressed while failing to explain why the current utterance reflects avoidance, hesitation, or a change in trust. A-MBER's explanation rubric gives the greatest combined weight to core correctness and key-point coverage, and separately measures rationale alignment. This structure makes the historical mechanism part of the target rather than treating a fluent affective label as a complete answer.

The qualitative example in Appendix D illustrates the distinction. The anchor is a question about APA citation formatting. In isolation, it admits a straightforward academic reading. The earlier description of a deadline problem as a minor time-management glitch, together with the current discussion of fatigue, supports the more specific interpretation that the academic question shields the student from uncomfortable personal scrutiny. The contrast lies in the use of the utterance within the interaction, not in the emotional valence of the words "APA citation formatting."

A correct explanation should connect the present shift to the relevant earlier pattern. It need not reproduce the reference wording. Conversely, a detailed story that invents an earlier event should lose credit under the unsupported-claims dimension even if it names the correct broad emotion. This is why the rubric includes both positive coverage criteria and a criterion for evidential discipline.

J.2 Memory Level and Reasoning Structure

Memory level and reasoning structure provide two views of task difficulty. Level 0 and nearby-context items act as local controls. Cross-session items require access beyond the current session. The highest memory level requires integrating multiple events, affective shifts, or relational development. A large context gain on these subsets is consistent with the intended historical dependence of the benchmark.

Reasoning structure distinguishes how available evidence must be used. A single-hop item may be settled by one salient earlier contrast. Multi-hop items require connecting more than one piece of history. Trajectory-based items require reconstructing a direction of change across interactions. Conflict-resolution items require comparing candidate interpretations against the evidence. Strong retrieval alone does not guarantee success on the latter three patterns because identifying a turn and using it in the right inferential relationship are different operations.

We report the largest historical-access improvements on multi-hop and trajectory-based items. This pattern supports using both context interventions and reasoning annotations in evaluation. It should not be read as a direct observation of the model's internal reasoning process: the benchmark observes answers, evidence predictions, and explanations under controlled input policies.

J.3 Robustness as Selective Use of History

Remembering more information is useful only when the remembered information supports the interpretation being evaluated. Pseudo-relevant histories test the tendency to mistake topical similarity for explanatory relevance. Insufficient-evidence items test whether the model acknowledges that the available record cannot support a confident conclusion. Missing or ambiguous delivery descriptions test dependence on cues beyond the literal wording of a dialogue turn.

These stress tests are related but distinct. A missing cue reduces the available evidence; an ambiguous cue allows competing readings; a pseudo-relevant event may actively distract from the relevant mechanism. Their reported scores should therefore be interpreted alongside the standard condition

rather than collapsed into a single notion of robustness. The main robustness table preserves the individual conditions for this reason.

The structured delivery descriptions retained in the trajectories should also be distinguished from naturally recorded audio or video. Our construction pipeline retains tone, pauses, and expressive style as annotations in the generated material. Consequently, modality-related results concern the availability and ambiguity of those supplied cues within this benchmark; they do not establish performance on a separately collected speech or vision dataset.

K Score Aggregation and Reproducibility Notes

K.1 Item Weighting and Paired Comparisons

The primary aggregate gives equal weight to each evaluated item. If A_t denotes the set of items scored under task family t, and the task sets form the scored partition, then

$$S^{(c)} = \sum_t \frac{|A_t|}{N} S^{(c)}_{A_t}. \tag{13}$$

This expression explains why the overall result need not equal the unweighted mean of the three task-family averages. Our item-level aggregation preserves benchmark composition, whereas an unweighted task macro-average answers a different question.

For a paired comparison, define $\delta_i = s_i^{(c')} - s_i^{(c)}$. A bootstrap replicate samples the same item indices for both context conditions:

$$I^{(b)} = (i_1^{(b)}, \ldots, i_N^{(b)}), \qquad \widehat{\Delta}^{(b)} = \frac{1}{N} \sum_{j=1}^{N} \delta_{i_j^{(b)}}. \tag{14}$$

The essential pairing is at the item level, so the difference is not confounded by evaluating one condition on easier items. The confidence intervals reported in the main text are inherited from the evaluation protocol in the main text.

Items also share trajectories. Item-level resampling measures uncertainty under the reported item-resampling protocol. It should not be interpreted as an independent estimate of generalization to unseen naturally occurring users or trajectories. Such a claim would require an evaluation and resampling design matched to that unit of generalization.

K.2 Interpreting Human Verification and Judge Validation

The human quality audit and the judge-validation study answer different questions. The audit examines benchmark items: whether a question is answerable, its gold answer supported, its evidence sufficient, and its metadata valid. The judge study examines scores assigned to model-generated explanations. A benchmark can have coherent items while an automatic judge remains imperfect, so the two checks should not be treated as substitutes.

We report 410 audited items, with 383 retained, 21 requiring minor revision, and six removed or rewritten. These counts describe the audited subset. We also report 180 explanation outputs used for judge validation, with a Spearman correlation of 0.78 and a mean absolute score difference of 0.083. These statistics support reporting the judge as a calibrated evaluation component while retaining the limitations of automatic explanation scoring.

The insufficiency-handling criterion is conditional: the supplied judge prompt sets it to 1.0 when insufficiency is irrelevant to the item. The unsupported-claims score is oriented in the opposite way from a raw penalty: a larger value denotes better evidential discipline. These conventions matter when interpreting the weighted score and when implementing the JSON schema in Appendix G.

K.3 Interpreting the Experimental Record

The model tables, subset analyses, human-audit counts, and judge-validation statistics address complementary aspects of benchmark validity. The mathematical definitions above make the relationships

among context policies and scores explicit. Their purpose is to support interpretation of the reported evaluation rather than to posit an additional optimization objective.

The per-item logging protocol records the item and model identifiers, context condition, metadata, prediction, references, and task-specific score. Those records are the appropriate basis for checking aggregate consistency, alternative weighting schemes, and robustness under different resampling units. The distinction between the manuscript and the underlying records is especially relevant when interpreting rounded aggregate tables or comparing subsets with different item compositions.

L Detailed Reporting Protocol

All model outputs are logged at the item level, with one record per model, context condition, and benchmark item. Each record includes the item ID, model ID, context condition, task family, content type, memory level, reasoning structure, robustness tag, model prediction, gold answer, gold evidence, and score. All aggregate and stratified results are computed from these item-level records using the same scoring script. This logging format allows the results to be traced back to individual benchmark items and supports consistent comparisons across context conditions and model families.

For significance testing, we use paired bootstrap resampling over benchmark items. For each comparison, we resample items with replacement and recompute the mean score difference between two context conditions using the same sampled item set. We report bootstrap confidence intervals for the local-context versus full-history, full-history versus retrieved-history, and local-context versus retrieved-history comparisons. This paired design controls for item difficulty and tests whether performance changes are driven by consistent item-level gains rather than by a small number of unusually easy or hard examples.

M Extended Related Work and Benchmark Positioning

A-MBER lies at the intersection of conversational affect, historically grounded interpretation, and memory evaluation. This review distinguishes the capabilities studied in these literatures and explains their relationship to our benchmark. The cited systems and datasets are conceptual comparators, rather than additional baselines evaluated in our experiments.

M.1 Contextual Emotion Recognition

The dependence of emotion on conversational history predates large language models. Conversational Memory Networks and ICON model interactions between speakers and their evolving states (Hazarika et al., 2018b;a). HiGRU introduces a hierarchical recurrent view of utterances, while knowledge-enriched transformers incorporate conversational context and external knowledge (Jiao et al., 2019; Zhong et al., 2019). These approaches establish why a sentence-level representation alone can be inadequate for affect recognition.

Graph-based approaches make different structural assumptions. Relation-aware attention encodes relationships and relative positions, DAG-ERC represents directed dependencies, and DualGATs combines speaker-aware and discourse-oriented views (Ishiwatari et al., 2020; Shen et al., 2021; Zhang et al., 2023). CoMPM incorporates pretrained speaker memory (Lee & Lee, 2022). Adversarial contrastive objectives address robust representations (Hu et al., 2023). GoEmotions broadens resources for fine-grained textual affect (Demszky et al., 2020).

Work on emotion triggers asks whether prediction relies on events that evoke an emotion (Singh et al., 2024). A-MBER extends this concern to earlier interaction events that justify the current interpretation. Modeling context within a labeled conversation is therefore a relevant precedent, while the anchor-centered, cross-session evaluation target remains distinct.

M.2 Multimodal Affect and Missing Cues

Multimodal affect recognition combines information whose reliability and temporal roles may differ. Joint sentiment–emotion modeling studies related prediction targets (Akhtar et al., 2019), while MMGCN and COGMEN represent contextual interactions in multimodal emotion recognition (Hu et al., 2021b; Joshi et al., 2022). UniMSE explores a unified treatment of sentiment analysis and emotion recognition (Hu et al., 2022).

Layer-wise fusion with modality independence, correlation-aware fusion in MultiEMO, and cross-modality contextual refinement offer different ways to combine signals (Sun et al., 2023; Shi & Huang, 2023; Zhang & Li, 2023). TelME investigates teacher-led multimodal learning (Yun et al., 2024). Missing Modality Imagination Networks and prompt learning with missing modalities directly address incomplete inputs (Zhao et al., 2021; Guo et al., 2024).

A-MBER's missing and ambiguous modality tags share the concern that affective cues can be incomplete. However, our generated trajectories retain structured delivery descriptions, rather than constituting a new collection of recorded speech and video. These references motivate a diagnostic dimension; their multimodal benchmark results should not be equated with performance on our description-based cues.

M.3 Emotion Causes and Evidence Relationships

Joint emotion classification and cause detection makes the explanatory event a prediction target (Chen et al., 2018). Emotion-cause pair extraction removes the requirement that the emotion be supplied in advance (Xia & Ding, 2019). This is an important precursor to evaluating an interpretation together with supporting evidence rather than considering classification sufficient on its own.

Subsequent approaches model the pair structure through two-dimensional representations (Ding et al., 2020). Pair-based joint encoding and constrained learning study relationships between candidate emotion and cause spans (Liu et al., 2022; Feng et al., 2023). Conversational emotion-cause extraction adapts this problem to dialogue-specific relationships (Jeong & Bak, 2023).

A-MBER shares the requirement to ground affect in observable events, but its evidence can concern a recurring pattern, an earlier repair attempt, or a contrast across sessions. The question is not always which clause states a cause. It can instead be why the current utterance has a particular relational meaning given a sequence of interactions. Its retrieval and explanation tasks preserve the distinction between locating evidence and synthesizing its implications.

M.4 Empathy as Understanding and Response Generation

EmpatheticDialogues provides a benchmark for recognizing a conversational partner's feelings and responding appropriately (Rashkin et al., 2019). MoEL and MIME explore emotion-conditioned response generation (Lin et al., 2019; Majumder et al., 2020). Persona-based empathetic models connect individual information to response behavior (Zhong et al., 2020). Computational work on empathy in text-based mental-health support distinguishes how a response expresses understanding (Sharma et al., 2020).

More specific approaches study perspective-taking around emotion causes, empathetic intents, and feature transitions across utterances (Kim et al., 2021; Chen et al., 2022; Kim et al., 2022). CASE connects cognition and affection, while associative-memory models iteratively connect relevant dialogue information (Zhou et al., 2023b; Yang et al., 2024). Empathetic alignment makes perspective-taking explicit (Yang & Jurgens, 2024). Context-insensitive empathy identification illustrates why apparently empathetic output is not conclusive evidence of contextual understanding (Lee et al., 2023).

A-MBER evaluates an upstream capability: interpreting the current state with appropriate historical justification. It does not assume that a high interpretation score is sufficient for effective support. Conversely, a response that sounds sympathetic need not demonstrate that the model understood the relevant history. This motivates evaluating interpretation, evidence, and explanation directly.

M.5 Emotional Support and Strategy over Time

ESConv establishes emotional support conversation as a structured dialogue problem (Liu et al., 2021). MISC, lookahead strategy planning, and knowledge-enhanced mixed-initiative dialogue examine how strategies and user states interact over multiple turns (Tu et al., 2022; Cheng et al., 2022; Deng et al., 2023). Positive-emotion elicitation and counseling agents study supportive actions and conversational development (Zhou et al., 2023a; Mishra et al., 2023).

LLM-based work explores synthetic self-chats, interpretable support reasoning, preference bias, and the relevance of responses across a support conversation (Zheng et al., 2024; Zhang et al., 2024b; Kang et al., 2024; Li et al., 2024). ESC-Eval focuses on evaluating emotional-support conversations (Zhao et al., 2024a). These studies show that support quality involves more than assigning a label or generating a single sympathetic sentence.

Their connection to A-MBER is the need to track how previous support attempts affect a user's present stance. An earlier reassurance may have helped, failed, or been experienced as dismissive. The same current wording can therefore require different interpretations. A-MBER makes this historical dependence observable while leaving intervention effectiveness outside the scope of the reported scores.

M.6 Personalized Dialogue and Persistent Memory

Long-term open-domain conversation and persona memory move beyond isolated exchanges (Xu et al., 2022a;b). Personal-memory-enhanced knowledge-grounded dialogue, history refinement, and split-memory response generation study how stored information influences subsequent responses (Fu et al., 2022; Zhong et al., 2022; Wu et al., 2021). Persona commonsense and retrieval-enhanced personalized generation address the interpretation and selection of user-related information (Gao et al., 2023; Liu et al., 2023a).

Recent approaches include commonsense-augmented memory management, editable memory graphs, and coordinated short- and long-term memory for personalized assistants (Kim et al., 2024; Wang et al., 2024d; Zhang et al., 2024a). LongMem separates a memory encoder from an adaptive retrieval-and-reading network (Wang et al., 2023a). These works span explicit memory representations, retrieval policies, and model adaptation.

A-MBER isolates context access to make comparisons interpretable. Its full-history and retrieved-history settings do not exhaust memory architectures. They provide controlled inputs on which historically grounded affective interpretation can be measured. Future evaluations can supply those inputs through more elaborate writing, updating, or forgetting mechanisms.

M.7 Long Context: Access, Retrieval, and Integration

Lost in the Middle demonstrates that larger input windows do not ensure equally effective information use at different positions (Liu et al., 2024). LongBench, L-Eval, and Ada-LEval provide complementary multi-task and length-sensitive evaluation resources (Bai et al., 2024; An et al., 2024; Wang et al., 2024a). Long-context scaling and prompt compression study access and the concentration of useful information (Xiong et al., 2024; Jiang et al., 2024). Unlimiformer uses a nearest-neighbor index to extend attention over long inputs (Bertsch et al., 2023).

Several benchmarks question whether locating a single passage adequately tests long-context understanding. Summary of a Haystack examines aggregation, extended multi-document QA emphasizes coverage, and NoCha tests claims about book-length narratives (Laban et al., 2024; Wang et al., 2024c; Karpinska et al., 2024). Work distinguishing retrieval from difficult long-context tasks makes the integration demand explicit (Goldman et al., 2024).

A-MBER follows this distinction in an interactional setting. Remembering an event is useful only if the model relates it correctly to the anchor. Multi-hop and trajectory-based subsets complement measures of input capacity and isolated retrieval. Chronological order, changes in stance, and the distinction between salience and relevance remain meaningful even when every turn fits into the context window.

M.8 Retrieval and Evidence Quality

Sentence-BERT supports efficient semantic comparison, and dense passage retrieval makes learned query–passage representations central to open-domain QA (Reimers & Gurevych, 2019; Karpukhin et al., 2020). RocketQA and corpus-aware pretraining further study retrieval representations (Qu et al., 2021; Gao & Callan, 2022). These are precedents for embedding-based selection, although A-MBER's experimental retriever is BAAI/bge-m3 rather than one of these earlier models.

REPLUG augments a black-box language model with retrieved text, Adaptive-RAG adjusts retrieval to question complexity, and LongRAG studies long-context QA (Shi et al., 2024; Jeong et al., 2024; Zhao et al., 2024b). Imperfect retrieval can supply unanswerable, adversarial, or conflicting material (Park & Lee, 2024). RAGTruth evaluates unsupported claims even when retrieval is available (Niu et al., 2024). ARES and RAGAs distinguish retrieval relevance from answer quality and groundedness (Saad-Falcon et al., 2024; Es et al., 2024).

These distinctions motivate separate evidence-retrieval and explanation scores. A turn may be semantically similar to the anchor yet provide little explanatory support. A correct evidence set also does not ensure that an answer uses it without overreach. Pseudo-relevant-history and insufficient-evidence conditions probe these failure modes within personal interaction histories.

M.9 Judging Explanations and Testing Reliability

BLEURT and SummEval motivate checking automatic scores against human judgments (Sellam et al., 2020; Fabbri et al., 2021). G-Eval explores LLM-based text assessment, while MT-Bench and Chatbot Arena examine the usefulness and biases of LLM judges (Liu et al., 2023b; Zheng et al., 2023). FActScore, SelfCheckGPT, and HaluEval study complementary aspects of factual precision and hallucination (Min et al., 2023; Manakul et al., 2023; Li et al., 2023). Correctness and faithfulness in instruction-following QA, and TofuEval's dialogue-focused analysis, distinguish a plausible answer from a grounded one (Adlakha et al., 2024; Tang et al., 2024).

ERASER benchmarks rationalized predictions, while faithfulness tests and analyses of self-consistency caution against treating an explanation as a transparent record of internal computation (DeYoung et al., 2020; Atanasova et al., 2023; Parcalabescu & Frank, 2024). Chain-of-thought prompting can improve reasoning, but generated reasoning text remains an observable output rather than direct access to internal mechanisms (Wei et al., 2022). A-MBER evaluates support from the dialogue and reference rubric, without claiming to verify internal reasoning traces.

CheckList and Dynabench motivate evaluation beyond a single static accuracy number (Ribeiro et al., 2020; Kiela et al., 2021). A-MBER's metadata and controlled context comparisons serve this diagnostic purpose. Human judge validation provides an empirical calibration check, while automatic scoring limitations remain relevant.

M.10 Social, Causal, and Temporal Commonsense

Event2Mind studies intents and reactions, while work on naive psychology annotates characters' mental states in narratives (Rashkin et al., 2018b;a). Social IQa targets social situations, and COMET learns to generate commonsense knowledge (Sap et al., 2019; Bosselut et al., 2019). GLUCOSE grounds implicit causal knowledge in narrative context (Mostafazadeh et al., 2020). TimeDial addresses temporal commonsense in dialogue, and ConvoSense expands conversational commonsense inferences (Qin et al., 2021; Finch & Choi, 2024).

These resources explain why affective memory cannot be reduced to matching repeated words. Interpretation can depend on an unstated intention, an expected reaction, or event order. However, general commonsense cannot substitute for the particular user's history. A typical reaction to disappointment does not establish that this user is disappointed at the current anchor. A-MBER requires connecting general interpretive knowledge to the available trajectory evidence.

M.11 Synthetic Construction, Documentation, and Scope

Roundtrip consistency and synthetic QA generation establish precedents for generating candidates and filtering them for answerability (Alberti et al., 2019; Puri et al., 2020). Self-Instruct studies generated instructions, while Self-Refine explores iterative feedback and revision (Wang et al., 2023b; Madaan et al., 2023). These precedents motivate distinguishing candidates from curated benchmark items; they do not remove the need for trajectory-level review and item-level human auditing.

Data statements emphasize documenting the collection process and represented settings (Bender & Friedman, 2018). Emotion-analysis ethics makes the risks and assumptions of affect inference explicit (Mohammad, 2022). XiaoIce illustrates sustained social interaction, while gendered emotion attribution in LLMs cautions against assuming judgments are neutral across people (Zhou et al., 2020; Plaza-del Arco et al., 2024).

A-MBER evaluates grounded interpretation in constructed trajectories. It does not diagnose mental health, establish intervention effectiveness, or measure every aspect of a deployed companion. Its contribution is the controlled relationship between present interpretation, historical evidence, and uncertainty, assessed through complementary tasks and context policies.