NaturalMem: A Benchmark for Memory-Driven Dialogue in Large Language Models

Anonymous ACL submission

Abstract

Most existing benchmarks for evaluating the memory of large language models (LLMs) rely on explicit recall-style question answering, where memory is directly queried and answered. However, such memory answering settings diverge from real world human–AI interaction, in which memory is rarely triggered explicitly and instead manifests implicitly by shaping dialogue generation. We introduce **NaturalMem**, a natural dialogue-based benchmark for evaluating **memory-driven dialogue**, where memory influences responses and speaking style without explicit recall prompts. **NaturalMem** constructs multi-turn dialogues based on fictional character prototypes, excluding identifiable names or show- specific entities. Each character is associated with personal facts, category information, and a target speaking style, and is evaluated across multiple dialogue sessions to assess fact retention and style consistency. Dialogue data is created through an LLM-assisted and human-curated pipeline. Experiments show that state-of-the-art agents still struggle to retain personal facts and maintain stable speaking styles in memory-driven dialogue settings. **NaturalMem** provides a realistic and diagnostic framework for evaluating memory in LLMs.

1 Introduction

Large language models (LLMs) are increasingly deployed as long-term conversational agents, personal assistants, and interactive systems that engage users over extended periods of time. In such settings, memory is a core capability: models are expected not only to process the immediate conversational context, but also to retain and consistently utilize user-specific information, personal facts, and behavioral patterns across interactions. Accordingly, evaluating the memory capabilities of LLMs and agent systems has emerged as an important research direction, leading to a growing number of benchmarks that assess how models store, retrieve, and use past information (Gao et al., 2025; Hu et al., 2025; Wu et al., 2025; Zhang et al., 2025).

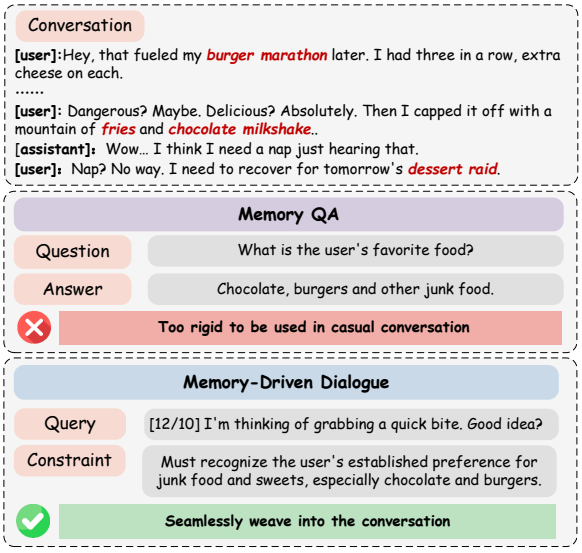

Figure 1: Transition from memory-based question answering to memory-driven dialogue.

Most existing memory benchmarks, however, conceptualize memory primarily as an object of explicit querying. Typical evaluations adopt recall-style question answering, where models are directly asked to restate previously provided information (e.g., "What did the user say earlier?") (Yang et al., 2018). While such **memory answering** paradigms are easy to quantify and provide clear evaluation targets, they rely on interaction patterns that are rare in natural human–AI conversations (Maharana et al., 2024; Wu et al., 2024; Zhong et al., 2024; Bai et al., 2024). In practice, users seldom probe a system's memory explicitly. Instead, memory is expected to operate implicitly, shaping responses in a way that is consistent with prior interactions, as illustrated in Figure 1.

Table 1: Statistics of long-form conversational benchmarks from LongMemEval (Wu et al., 2024). These benchmarks differ in domain, scale, and interaction setting. DialSim is constructed from existing TV show transcripts, while NaturalMem consists of fictional yet naturalistic dialogues generated using real TV characters, specifically designed for controlled evaluation of long-term conversational memory.

Benchmark	Domain	Sess.	Ques.	Tokens / Conv.	Conversation Type
MSC(Xu et al., 2022)	Open-Domain	5k	–	1k	Human–Human
MemoryBank(Zhong et al., 2024)	Personal	300	194	5k	Human–AI
PerLTQA(Du et al., 2024)	Personal	4k	8,593	1M	Human–AI
LoCoMo(Maharana et al., 2024)	Personal	1k	7,512	10k	Human–Human
DialSim(Kim et al., 2024)	TV Shows	1k–2k	1M	350k	Human–Human
LongMemEval(Wu et al., 2024)	Personal	50k	500	115k, 1.5M	Human–AI
PersonaMem-V2(Jiang et al., 2025b)	Personal	–	5000	32k, 128k	Human–AI
NaturalMem (ours)	Fictional TV	27k	261	14k	Human–Human

This discrepancy reveals a fundamental gap between current benchmarks and realistic conversational use cases. In real-world dialogue, memory functions less as a mechanism for explicit recall and more as a behavioral driver of dialogue generation. A model demonstrates memory not by stating what it remembers, but by responding in ways that remain consistent with previously established facts, preferences, and interaction styles (Shuster et al., 2022). Importantly, this notion of memory-driven behavior differs from prompted or instruction-following personas. Rather than being repeatedly specified in the prompt, stylistic traits and personal characteristics are introduced implicitly through earlier interactions and must be retained and applied without explicit restatement.

To address this gap, we introduce **NaturalMem**, a benchmark designed to evaluate memory-driven dialogue in natural conversational settings. NaturalMem constructs dialogue scenarios based on characters drawn from television programs, whose coherent personal traits and stable speaking styles provide a realistic foundation for modeling long-term memory effects in dialogue. Each character is associated with a structured background specification, enabling systematic evaluation of both **factual consistency** and **stylistic stability** across interactions (Zhang et al., 2018). To prevent models from exploiting prior world knowledge, NaturalMem deliberately excludes character names and show-specific entities from the dialogue text, ensuring that successful performance requires genuine memory usage rather than surface-level recognition.

We evaluate a range of large language models and memory-augmented variants under this setting. Our results show that even models with strong explicit recall abilities often fail to consistently maintain personal facts and stable speaking styles when memory must be applied implicitly in natural dialogue. These findings suggest that recall-centered benchmarks substantially underestimate the challenges of memory in realistic conversational interactions. NaturalMem is not intended to replace existing long-term memory benchmarks such as LoCoMo or LongMemEval, which focus on explicit recall or comprehension over extended dialogue histories. Instead, it targets a complementary and underexplored regime: whether memory can behaviorally and implicitly drive dialogue generation in the absence of explicit memory queries.

2 Related Work

2.1 Memory Modules for LLMs

Research on memory in language models spans from early neural architectures with explicit read–write memory to recent system-level memory augmentation for large language models. Early work introduced external memory mechanisms, such as Memory Networks (Weston et al., 2014), Neural Turing Machines (Graves et al., 2014), and Differentiable Neural Computers (Graves et al., 2016), with the goal of enabling models to store and retrieve information beyond the immediate input context.

Building on these foundations, recent systems incorporate practical memory components into LLM-based agents, including Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025), MemoryOS (Kang et al., 2025), and MemOS (Li et al., 2025). These approaches store user-specific facts or interaction histories and retrieve them to condition future generation, reflecting a broader trend toward treating memory as a first-class component in LLM applications rather than relying solely on extended context windows.

Despite their promise, such memory-augmented

systems are often evaluated using task-specific or ad hoc protocols. Consequently, it remains unclear whether the retrieved memory meaningfully influences natural dialogue behavior, or merely serves as an external lookup mechanism, highlighting the need for systematic and realistic evaluation settings.

2.2 Memory Evaluation Benchmarks

A number of benchmarks have been proposed to evaluate memory and long-context capabilities of LLMs, most of which adopt recall-oriented paradigms. In these settings, memory is explicitly queried through question answering, as exemplified by MemoryBank (Zhong et al., 2024) and LongBench (Bai et al., 2024). More recent benchmarks such as LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2024) extend this paradigm to longer and multi-session dialogues, yet still primarily rely on explicit probing.

Beyond evaluation protocols, existing long-form conversational memory benchmarks also vary widely in their domains, scales, and interaction settings. As summarized in Table 1, prior datasets differ in whether conversations are Human–Human or Human–AI and the typical conversation length. Consequently, these benchmarks predominantly measure memory answering, rather than whether memory implicitly shapes dialogue generation in realistic conversational settings.

2.3 Implicit Memory and Style Consistency

Beyond factual recall, prior work has emphasized persona and stylistic consistency as important aspects of dialogue systems. Persona-based benchmarks such as Persona-Chat (Zhang et al., 2018) show that maintaining stable personal attributes improves conversational coherence. More recent efforts further explore related dimensions: LikeBench (Awsafur Rahman et al., 2025) evaluates likability and adaptability in multi-session personalized interactions, while PersonaMem (Jiang et al., 2025b,a) focuses on modeling implicit personas and evolving user profiles over time.

These studies underscore that memory in dialogue systems is often expressed implicitly, through consistent behavior and linguistic style rather than explicit recollection. However, existing benchmarks typically evaluate persona or style in isolation, without jointly assessing factual consistency and stylistic continuity in natural, multi-session dialogue.

NaturalMem differs from prompted or instruction-following persona settings commonly studied in prior work. Prompted personas rely on explicit conditioning, where stylistic attributes are repeatedly specified or reinforced in the prompt. In contrast, the style consistency evaluated in NaturalMem is memory-driven: stylistic traits are introduced implicitly through earlier interactions and must be retained and applied in later sessions without restatement.

Maintaining such behavior therefore requires models to utilize long-term contextual information across sessions rather than responding to local instructions. In this sense, style consistency in NaturalMem constitutes a form of implicit memory, manifested behaviorally through stable linguistic patterns, tone, and interactional preferences over time, even in the presence of distractors.

3 Reverse-Guided Dataset Construction

Fully automatic dataset construction using large language models alone (Jiang et al., 2025b), corresponding to LLM draft generation without human intervention (Figure 2, Stage 2), is often insufficient for natural conversational memory evaluation. In practice, LLM-generated drafts frequently violate de-identification constraints, exhibit repetitive question patterns, and fail to support deep, distributed memory dependencies required by downstream evaluation (Figure 2, Stage 4).

To address these limitations, we propose **Reverse-Guided Dataset Construction (RGDC)**, a human-intensive (Long et al., 2024; Maharana et al., 2024) and iterative pipeline that treats LLM outputs as intermediate drafts rather than final artifacts. As illustrated in Figure 2, RGDC integrates structured human curation and feedback (Figure 2, Stage 3) to iteratively correct generation failures and guide constrained regeneration.

3.1 LLM-Based Draft Dialogue Generation

As shown in Figure 2 (Stage 2), dialogues are first generated by an LLM conditioned on abstract role descriptions and multi-session conversational contexts. Explicit de-identification constraints are imposed to discourage character names and other identity-revealing references, ensuring that correct answers must be supported by prior conversational evidence rather than external knowledge.

Despite these constraints, drafts produced at this stage often contain identity leakage, stereotypical

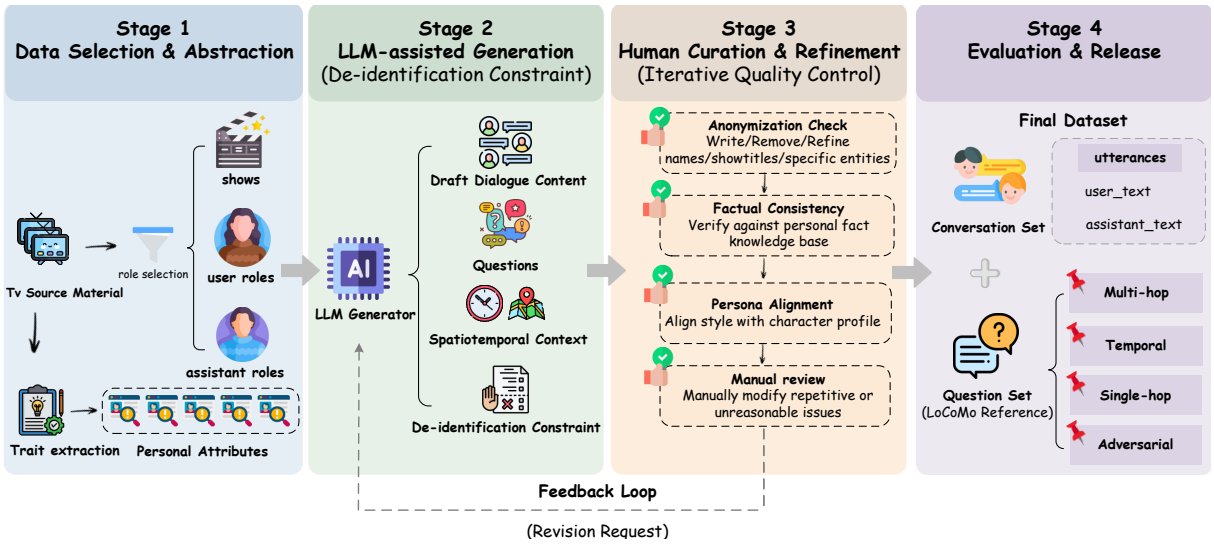

Figure 2: Reverse-Guided Dataset Construction (RGDC), an iterative human-in-the-loop pipeline. The process consists of four stages: (S1) data selection and abstraction, (S2) LLM-assisted draft generation under de-identification constraints, (S3) human curation and refinement with structured quality control, and (S4) evaluation-oriented dataset assembly. Common failure modes in Stage 2 are addressed through iterative feedback and constrained revision in Stage 3, forming a closed loop that targets the reasoning requirements of Stage 4.

expressions, and shallow memory cues. All generated dialogues are therefore treated as provisional drafts and forwarded to human curation (Figure 2, Stage 3).

3.2 Human Curation and Structural Enforcement

To correct systematic failures from LLM drafting (Figure 2, Stage 2), all dialogues undergo manual inspection during human curation (Figure 2, Stage 3).This step focuses on enforcing structural constraints that are difficult to guarantee through prompting alone, particularly the removal of identity leakage and shortcut signals that could enable reliance on pretraining knowledge.

Edits are performed through targeted manual rewriting rather than automatic deletion, as naive removal often disrupts discourse coherence (Wu et al., 2024).

3.3 Reverse-Guided Question Construction

Evaluation questions are explicitly decoupled from initial dialogue generation (Figure 2) and constructed in a reverse-guided manner. Questions generated directly by LLMs tend to be repetitive and biased toward simple memory patterns.

Human annotators therefore design evaluation questions first (Figure 2, Stage 4), targeting diverse reasoning requirements, including multi-hop, temporal, and adversarial memory dependencies. Dialogue content is subsequently revised or regenerated so that answers are naturally grounded in prior conversations, forming an explicit feedback loop (Figure 2).

3.4 Iterative Human-in-the-Loop Refinement

Overall, RGDC operates as an iterative human-in-the-loop process rather than a linear pipeline, as explicitly illustrated by the feedback loop in Figure 2. Human verification and editing are carried out by multiple annotators drawn from the same research group, all of whom are familiar with the construction guidelines and independently review the generated dialogues. Revisions are cross-checked by at least one additional annotator to reduce individual bias and ensure consistency.

Dialogues often undergo multiple rounds of inspection, manual editing, constrained regeneration, and re-verification across Stage 2 and Stage 3. Correcting one issue (e.g., removing an identity cue) may introduce others (e.g., reduced coherence or loss of implicitness), requiring further iteration. Question sets are also repeatedly reviewed to eliminate redundancy; when duplicate or overly similar questions are identified, annotators manually revise or replace them, followed by corresponding dialogue adjustments. This iterative refinement continues until dialogues satisfy constraints on naturalness, memory dependency, and robustness, or are discarded.

4 Evaluation Benchmark

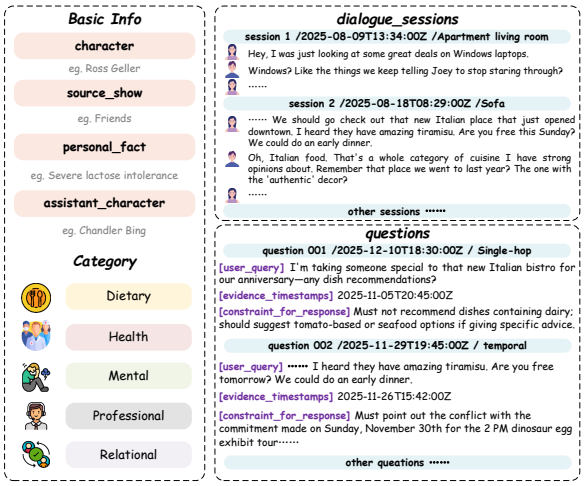

Figure 3: Schema of our constructed dataset. Each instance comprises a user character persona derived from a TV show, an assistant character style, multiple anonymized dialogue sessions, and a set of evaluation questions with grounded factual constraints.

Dataset overview. Based on the multi-session dialogues constructed in §3, we introduce an evaluation benchmark designed to measure long-term conversational memory under identity-anonymized settings. The dataset comprises **73** unique `source_show`, **90** distinct `character` personas, and **89** `assistant_character` styles. In total, it contains **2,774** `dialogue_sessions` with **43,492** utterances, amounting to approximately **1.3M** tokens. The structure of a single instance is illustrated in Figure 3. We further annotate **261** evaluation questions spanning diverse long-term memory reasoning requirements.

Crucially, all dialogues are identity-anonymized: character names and show-specific entities are removed from the text. This design ensures that successful performance cannot be achieved through entity recognition or prior world knowledge, but instead requires models to retain and utilize dialogue-grounded memory across sessions.

Personal fact categories. To characterize the content of conversational memory, we categorize personal facts appearing in dialogues into five types: Dietary, Health, Mental, Professional, and Relational. These categories reflect common forms of personal information naturally exchanged in everyday conversations and are designed to capture diverse memory demands.

Different categories place distinct requirements on long-term memory. For example, professional routines often involve stable, recurring facts, while mental or relational states may evolve gradually across sessions. By structuring memory targets along these dimensions, NaturalMem enables more fine-grained analysis of how models retain and utilize different types of personal information over time.

Question categories. Following prior work on multi-session memory evaluation (Maharana et al., 2024), we categorize evaluation questions into four types: (1) **Single-hop** questions depend on a single memory point, typically localized within one session; (2) **Multi-hop** questions require integrating multiple historical facts distributed across sessions; (3) **Temporal** questions involve time ordering, state transitions, or before/after relations that cannot be resolved from a single isolated mention; (4) **Adversarial** questions include misleading, conflicting, or underspecified cues, testing whether models can robustly preserve grounded facts and reject unsupported inferences.

This taxonomy is designed to reflect realistic conversational memory usage while preventing shortcut answering via character identity, show-specific knowledge, or external world facts. Unlike some prior benchmarks, we do not introduce a separate open-domain question category. This choice is intentional. In our identity-anonymized setting, every evaluation question already requires open-ended natural language generation grounded in dialogue history, rather than selecting or extracting a short factual span. Models must integrate recalled information with open-domain reasoning to produce a coherent response that satisfies the specified constraints. Consequently, open-domain reasoning is not treated as an isolated category, but is instead an inherent property of all question types in NaturalMem.

4.1 Task 1: Factual Memory Consistency

A conversational agent is expected to recall dialogue-grounded personal facts introduced implicitly over long histories and to maintain factual consistency in subsequent sessions. We formulate **Fact** as an accuracy-based evaluation. For each query, the agent generates a response that must satisfy a set of factual requirements encoded in `constraint_for_response`. A response is marked as correct if and only if it satisfies all specified constraints and does not contradict any

Table 2: Performance of base LLMs across different question categories and overall. Fact is reported as accuracy (%), Style as mean Likert score (1–5), and Latency as average wall-clock generation time (seconds) per query. All evaluations are conducted using DeepSeek-V3.2 as the unified LLM-as-a-Judge.

Model	Single-hop		Multi-hop		Temporal		Adversarial		Overall		Latency (s)
	Fact	Style	Fact	Style	Fact	Style	Fact	Style	Fact	Style	
DeepSeek-V3.2	75.6	4.07	64.9	4.23	**39.0**	4.06	54.5	4.01	55.6	4.08	3.37
Doubao-1.5-Pro-32k	77.8	4.13	**75.4**	4.21	30.5	4.11	**68.8**	4.09	**59.8**	4.13	6.07
Qwen3-235b-a22b	77.3	4.14	65.5	**4.55**	29.1	4.35	58.7	**4.40**	54.2	4.37	11.67
Qwen3-32b	79.5	4.21	56.4	4.53	30.4	4.28	50.7	4.15	50.6	4.28	6.99
Qwen3-30b-a3b	79.5	**4.50**	70.9	**4.55**	27.8	**4.38**	52.0	4.37	53.4	**4.44**	4.55
GPT-4o-mini	68.9	3.49	59.6	3.91	24.4	3.98	41.6	3.88	44.8	3.85	3.95
GPT-4.1-mini	**82.2**	3.49	66.7	3.79	**39.0**	3.68	42.9	3.92	53.6	3.74	**2.95**

dialogue-grounded facts; otherwise, it is marked incorrect. We report overall **Accuracy**, as well as accuracy broken down by question category.

LLM-as-a-Judge. We adopt an LLM-as-a-Judge protocol to determine factual correctness. Automatic n-gram overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and token-level F1 are ill-suited for this task, as valid answers may differ substantially in surface form while preserving identical factual content, and high lexical overlap may still conceal subtle contradictions. The judge model is instructed to verify constraint satisfaction based on the dialogue history and to output a binary correctness decision, along with a brief rationale for auditing purposes.

4.2 Task 2: Assistant Style Consistency

Beyond factual memory, natural long-term interaction also requires stable stylistic behavior. In the **Style** task, the assistant is expected to respond in a manner consistent with a target assistant style described by `assistant_character`, across multiple sessions and in the presence of distractors. Notably, the style specification is introduced implicitly through earlier interactions and is not restated in the evaluation prompt.

We evaluate style consistency using an LLM-as-a-Judge protocol that compares the generated response against the target style specification.

Scoring (1–5). Each response is rated on a five-point Likert scale (Likert, 1932): **1** = clearly off-style with strong drift; **2** = mostly off-style with occasional matching cues; **3** = partially consistent but unstable or generic; **4** = largely consistent with minor deviations; **5** = highly consistent and characterful throughout.

LLM-as-a-Judge. We use LLM-based evaluation because stylistic fidelity is inherently non-lexical and cannot be reliably captured by overlap-based metrics. BLEU, ROUGE, and F1 reward surface copying and penalize legitimate stylistic variation, while failing to measure higher-level signals such as tone, humor, formality, or verbosity specified in `assistant_character`. The judge is instructed to assess stylistic faithfulness independently of factual correctness.

4.3 Task 3: Answering Latency under Memory Retrieval

Memory-augmented agents may incur additional overhead due to retrieval, re-ranking, and prompt construction. We therefore measure end-to-end response time per query under a standardized inference setting. Latency statistics, including mean values, are reported to characterize efficiency–accuracy trade-offs across models.

5 Experiment Setup

All experiments are conducted under a unified text-only setting. Models are evaluated on anonymized multi-session dialogues without access to character identities, show-specific entities, or any external metadata. All systems are tested on the same dialogue histories and evaluation questions described in §4. To ensure comparability across models and tasks, all evaluations are standardized using a single fixed evaluation model.

Models. We evaluate both base large language models and memory-augmented agents to assess long-term conversational memory under diverse modeling paradigms. As base models, we include a representative set of state-of-the-art LLMs spanning different architectures, training regimes, and parameter scales: **DeepSeek-V3.2**, **Doubao-1.5-Pro-32k-250115**, **GPT-4o-mini**, **GPT-4.1-mini**, as well as three models from the **Qwen3** family:

Table 3: Performance of memory-augmented agents across different question categories and overall. Fact is reported as accuracy (%), Style as mean Likert score (1–5), and Latency as average wall-clock generation time (seconds) per query. All evaluations are conducted using DeepSeek-V3.2 as the unified LLM-as-a-Judge.

Method	Single-hop		Multi-hop		Temporal		Adversarial		Overall		Latency (s)
	Fact	Style	Fact	Style	Fact	Style	Fact	Style	Fact	Style	
GPT-4o-mini											
Mem0	40.0	2.38	47.4	2.33	18.3	2.46	6.5	2.48	24.9	2.43	4.76
A-Mem	48.9	2.93	63.2	2.96	**32.9**	3.24	16.9	3.08	37.5	3.08	3.61
MemoryOS	64.4	3.33	**70.2**	3.39	30.5	3.24	**51.9**	3.38	**51.3**	3.33	**2.39**
MemOS	46.7	3.07	59.6	3.09	11.0	3.15	3.9	3.34	25.7	3.17	3.58
Full-context	**68.9**	**3.49**	59.6	**3.91**	24.4	**3.98**	41.6	**3.88**	44.8	**3.85**	3.95
GPT-4.1-mini											
Mem0	37.8	2.36	42.1	2.21	23.2	2.52	6.5	2.62	24.9	2.46	4.38
A-Mem	64.4	3.00	71.9	3.16	36.6	3.44	28.6	3.40	46.7	3.29	2.91
MemoryOS	64.4	**3.82**	**73.7**	3.77	31.7	3.51	**50.6**	3.79	52.1	3.70	**2.23**
MemOS	48.9	3.29	70.2	3.44	13.4	3.51	6.5	3.65	29.9	3.50	4.36
Full-context	**82.2**	3.49	66.7	**3.79**	**39.0**	**3.68**	42.9	**3.92**	**53.6**	**3.74**	2.95

Qwen3-235B-A22B, **Qwen3-32B**, and **Qwen3-30B-A3B**. This selection covers both dense and mixture-of-experts architectures, and spans a wide range of model capacities, enabling a more comprehensive evaluation of memory behavior across scales.

To assess the effectiveness of explicit memory mechanisms beyond extended context, we additionally evaluate representative memory-augmented agents, including **Mem0**, **A-Mem**, **MemoryOS**, and **MemOS**. All memory-augmented agents are configured to store and retrieve dialogue histories according to their recommended settings. For fair comparison, identical dialogue inputs, prompts, and evaluation protocols are used across base LLMs and memory-augmented agents.

Evaluation Protocol. Factual memory consistency and assistant style consistency are evaluated using the LLM-as-a-Judge protocol defined in §4. Judgments for factual correctness and stylistic fidelity are performed independently. All models generate responses under the same prompting and decoding settings to control for generation variability.

Latency Measurement. To characterize efficiency trade-offs, we measure end-to-end response latency for all models when answering benchmark queries. Latency is recorded as wall-clock time per query under a standardized inference environment. For memory-augmented agents, the reported latency includes memory retrieval, re-ranking, and prompt construction overhead. We report average latency as well as percentile statistics to reflect practical deployment considerations. Detailed hardware configurations and runtime settings are provided in the Appendix A.1.

Evaluation Model. We use **DeepSeek-V3.2** as the primary evaluation model for all LLM-as-a-Judge judgments, including factual correctness verification and style scoring. Using a single evaluator ensures consistency and comparability across models and tasks.

To further verify robustness and diversity, we additionally employ **GPT-4o-mini** as an alternative judge model and perform manual spot-checks on randomly sampled responses. The overall trends and conclusions remain consistent across both automatic judges and human verification; detailed comparisons are provided in Appendix B.

6 Results and Analysis

We present the experimental results on NaturalMem from four perspectives: model scale effects, factual memory versus stylistic consistency, the behavior of memory-augmented agents, and efficiency–accuracy trade-offs. Additional analysis and discussion are provided in Appendix A.2.

6.1 Effect of Model Scale

Table 2 reports the performance of base LLMs on NaturalMem. Overall, we observe that larger or more capable models do not consistently achieve higher accuracy on this benchmark. In particular, improvements in single-hop factual questions do not reliably transfer to multi-hop, temporal, or adversarial settings.

This result contrasts with trends commonly observed in explicit recall or short-context reasoning benchmarks, indicating that NaturalMem captures a distinct aspect of conversational memory beyond raw model capacity.

6.2 Factual Memory vs. Style Consistency

As shown in Table 2, factual accuracy varies substantially across question categories, with the largest drops occurring in temporal and adversarial questions. In contrast, style consistency scores are relatively stable across models and categories.

This discrepancy suggests that factual consistency is more sensitive to long-term memory degradation than stylistic behavior.

Table 4: Accuracy (%) comparison between LoCoMo and NaturalMem. Δ denotes the accuracy difference.

Model	LoCoMo	NaturalMem	Δ
Mem0	66.3	24.9	41.4
A-Mem	62.1	37.5	24.6
MemoryOS	61.6	51.3	10.3
MemOS	73.3	25.7	47.6
Full-context	80.6	44.8	35.8

6.3 Memory-Augmented Agents

Results for memory-augmented agents are summarized in Table 3. Compared to base models, several agents achieve higher factual accuracy, demonstrating the benefit of explicit memory retrieval, but often at the cost of lower style consistency relative to the full-context baseline.

These results further suggest that memory selection itself may be a source of error, as retrieved content does not always align with the requirements of the current query. Among the evaluated systems, MemoryOS exhibits a more balanced performance across factual accuracy and style consistency, highlighting trade-offs in current agent-based memory designs.

6.4 Full-Context Baselines

Despite the absence of explicit memory retrieval, the full-context baseline achieves competitive and often higher factual accuracy than several memory-augmented agents (Table 3), indicating that imperfect retrieval and memory selection can partially offset the benefits of external memory mechanisms.

Notably, full-context models maintain substantially higher style consistency than most agent-based systems, suggesting that current agents primarily focus on fact-oriented memory extraction while stylistic cues are less effectively preserved.

However, full-context inference incurs higher computational costs, motivating the use of memory-augmented approaches in practical deployments.

6.5 Latency Analysis

Latency statistics are reported in Tables 2 and 3. For base models and full-context inference, latency increases with input length and is dominated by prefill and generation costs. Memory-augmented agents introduce additional overhead from memory retrieval and prompt construction, resulting in different efficiency–accuracy trade-offs.

6.6 Comparison with LoCoMo

Table 4 compares performance on LoCoMo and NaturalMem. All evaluated systems perform substantially worse on NaturalMem, confirming that it poses a more challenging evaluation setting.

Unlike LoCoMo, which explicitly probes memory through recall-style questions, NaturalMem evaluates whether memory can implicitly guide response generation in natural dialogue, as illustrated in Figure 1.

7 Conclusion and Future Work

We introduce **NaturalMem**, a benchmark for evaluating *memory-driven dialogue*, where memory implicitly guides factual consistency and speaking style rather than being explicitly queried. Our results show that both base LLMs and memory-augmented agents struggle under this setting, despite strong performance on recall-centric benchmarks, revealing limitations of existing memory evaluation protocols.

Several directions remain for future work. First, our current agent evaluation assumes that memory is always retrieved. In realistic deployments, agents must additionally decide when memory should be stored and whether retrieval is necessary for a given query. The ability to selectively trigger memory retrieval may not only improve memory utilization, but also reduce response latency by avoiding unnecessary retrieval and prompt construction. Second, extending NaturalMem to multimodal interactions may enable richer assessments of personalized and long-term consistency, moving toward a more comprehensive form of personalized Turing-style evaluation.

Limitations

NaturalMem has several **limitations**. First, the number of evaluation questions is relatively small, as question construction requires substantial human involvement to ensure naturalness and robust memory dependencies. Second, our agent evaluation does not model autonomous decisions about when to store or retrieve memory, focusing instead on memory utilization given retrieved content. Finally, while we employ LLM-as-a-Judge evaluation with manual spot-checking, automatic evaluation of memory-driven dialogue remains an open challenge.

References

Md Awsafur Rahman, Adam Gabrys, Doug Kang, Jingjing Sun, Tian Tan, and Ashwin Chandramouli. 2025. Likebench: Evaluating subjective likability in llms for personalization. *arXiv e-prints*, pages arXiv–2512.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. 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)*, pages 3119–3137.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. *arXiv preprint arXiv:2402.16288*.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, and 1 others. 2025. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, and 1 others. 2016. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025a. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

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

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. 2024. Dialsim: A real-time simulator for evaluating long-term dialogue understanding of conversational agents. *arXiv e-prints*, pages arXiv–2406.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, and 1 others. 2025. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung,

Moya Chen, Kushal Arora, Joshua Lane, and 1 others. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

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

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. 2025. From human memory to ai memory: A survey on memory mechanisms in the era of llms. *arXiv preprint arXiv:2504.15965*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. 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)*, pages 5180–5197.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47.

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

Appendix Overview

The appendix provides supplementary details to support reproducibility and further analysis of NaturalMem.

Section A describes the experimental setup, evaluation procedures, and additional analyses of memory behavior for both base models and memory-augmented agents.

Section B reports additional validation of the LLM-as-a-Judge protocol, including alternative judge models and human verification.

Section C provides the prompt templates used in experiments.

A Experiment Details and Analysis

A.1 Evaluation Procedures

We detail the evaluation procedures used for base models and memory-augmented agents to ensure reproducibility and fair comparison.

For base LLMs and the full-context baseline, the entire dialogue history is provided as input, with the final turn being the user query. Models are instructed to directly generate a response without any additional task-specific prompts or instructions.

For memory-augmented agents, dialogue histories are first processed to construct their internal memory representations according to each agent's recommended configuration. At inference time, retrieved memory content is concatenated with the user query using a minimal prompt template, and the model generates a response conditioned on the retrieved memory.

All models and agents are evaluated via API calls under comparable network conditions. Since all systems are invoked through API-based interfaces, network latency is consistent across models and can be safely ignored. Reported latency therefore primarily reflects model inference and memory retrieval overhead, ensuring the reliability of latency comparisons.

A.2 Additional Analysis of Results (Tables 2–3)

This subsection provides a detailed, table-grounded analysis to complement the high-level observations in §6.

(1) Category-wise factual difficulty: Temporal and adversarial are the bottlenecks. Table 2 shows a consistent pattern across base LLMs: *Single-hop* factual accuracy is relatively high (e.g., 68.9–82.2), *Multi-hop* drops moderately (56.4–75.4), while *Temporal* exhibits the largest degradation (24.4–39.0), and *Adversarial* remains challenging (41.6–68.8). This explains why gains on single-hop do not reliably transfer to more realistic settings: NaturalMem's hardest categories require maintaining state transitions and rejecting misleading cues, which are precisely where models fail

most.

(2) Model scale is not predictive under memory-driven dialogue. Despite large model capacity, Qwen3-235B-A22B achieves 54.2 overall factual accuracy, comparable to DeepSeek-V3.2 (55.6) and below Doubao-1.5-Pro-32k (59.8), while incurring much higher latency (11.67s vs. 3.37s and 6.07s, respectively). This non-monotonicity suggests that memory-driven dialogue performance depends less on raw capacity and more on whether the model can robustly *select* and *apply* distributed dialogue evidence, especially for temporal/adversarial constraints.

(3) Fact vs. style: style scores are comparatively stable, but do not imply strong factual memory. Across base LLMs, style consistency is relatively high and clustered (roughly 3.74–4.55), even when factual accuracy differs substantially (44.8–59.8 overall). For example, GPT-4o-mini attains 3.85 style with 44.8 fact, whereas Doubao reaches 4.13 style with 59.8 fact. This indicates that producing a style-consistent response is easier than satisfying strict factual constraints: a response can be stylistically plausible while still violating a constraint (e.g., temporal ordering, exclusions).

(4) Agent systems: factual gains come with a large style drop, especially versus full-context. Table 3 reveals a clear trade-off for memory-augmented agents. Under GPT-4o-mini, MemoryOS improves overall factual accuracy to 51.3, surpassing the full-context baseline (44.8), while also reducing latency (2.39s vs. 3.95s). However, its style score drops to 3.33 compared to full-context 3.85. A similar pattern holds for A-Mem (fact 37.5; style 3.08) and Mem0 (fact 24.9; style 2.43), both far below full-context style. This supports the hypothesis that current agent memory pipelines are primarily *fact-centric*: they retrieve and inject factual summaries well enough to help constraint satisfaction, but do not preserve the implicit stylistic cues accumulated across sessions, leading to style drift.

(5) Memory selection is a measurable error source. The agent results also suggest that retrieval *relevance/selection* is a key limitation. For instance, MemOS under GPT-4o-mini attains reasonable style (3.17) but very low factual accuracy (25.7), with catastrophic performance on adversarial (3.9) and temporal (11.0). Such patterns are consistent with retrieving incomplete or misaligned memories: the generation remains fluent (and sometimes stylistically acceptable) while failing hard on constraint satisfaction.

(6) Full-context remains a strong upper bound for style, and often for facts. Full-context generally provides higher style consistency than agent systems. For GPT-4o-mini, full-context style is 3.85, above MemoryOS 3.33 and A-Mem 3.08; for GPT-4.1-mini, full-context is 3.74, close to MemoryOS 3.70 and above A-Mem 3.29. On factual accuracy, full-context is also competitive: it reaches 53.6 on GPT-4.1-mini, slightly higher than MemoryOS 52.1, indicating that retrieval benefits can be offset by imperfect memory selection and compression.

(7) Latency trade-offs: retrieval can be faster than full-context, but not always. Latency profiles differ substantially. On GPT-4o-mini, MemoryOS reduces latency (2.39s) below full-context (3.95s) while improving facts, but Mem0 increases latency (4.76s) while performing much worse (24.9). On GPT-4.1-mini, MemoryOS is again faster (2.23s) than full-context (2.95s). These results suggest that retrieval-based pipelines can reduce end-to-end latency when memory selection is accurate and retrieval overhead is controlled; otherwise, retrieval adds cost without improving correctness. This motivates future work on *adaptive retrieval triggering* (deciding when retrieval is necessary), which could further reduce latency by avoiding unnecessary retrieval and prompt construction.

B Judge Model Robustness

B.1 Robustness to the Choice of LLM-as-a-Judge

In addition to using **DeepSeek-V3.2** as the primary LLM-as-a-Judge throughout our main experiments, we evaluate the robustness of factual correctness assessment by replacing the judge model with **GPT-4o-mini**. Crucially, all agent responses are kept identical across settings; only the evaluator is changed.

Table 5 reports the overall factual consistency of memory-augmented agents under the two evaluators. Although absolute accuracy values differ across judges, the relative performance trends remain stable. In particular, **MemoryOS** and **Full-context** consistently achieve the highest factual accuracy under both judges, whereas lightweight

Table 5: Overall factual consistency (%) of memory-augmented agents under different LLM-as-a-Judges. Agent responses are identical across settings; only the evaluator is changed. This table is included in the appendix as a robustness check.

Method	DeepSeek-V3.2 Judge	GPT-4o-mini Judge
Mem0	24.9	29.5
A-Mem	46.7	51.7
MemoryOS	52.1	**59.4**
MemOS	29.9	30.7
Full-context	**53.6**	56.7

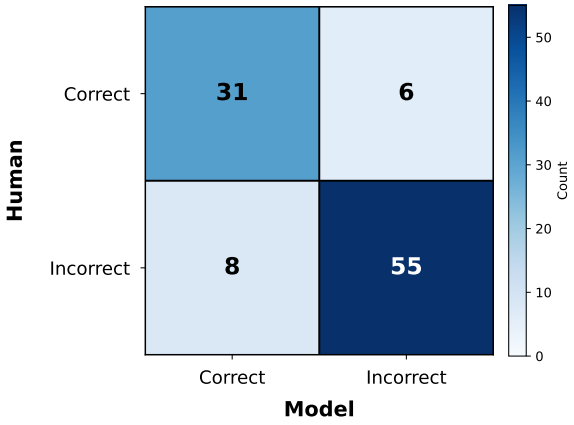

Figure 4: Confusion matrix comparing human annotations and DeepSeek-V3.2 factual judgments on 100 randomly sampled responses generated by MemoryOS. Diagonal entries indicate agreement between human evaluators and the LLM-based judge, while off-diagonal entries correspond to disagreement cases.

memory mechanisms such as **Mem0** and **MemOS** remain substantially weaker. This indicates that our main conclusions are not sensitive to the choice of LLM-based evaluator.

We further observe that GPT-4o-mini generally assigns higher factual correctness scores than DeepSeek-V3.2. Since the evaluated agent responses are identical, this discrepancy reflects differences in judgment strictness and calibration rather than changes in agent behavior. Such systematic shifts are expected when switching between large language models with different evaluation granularity and error tolerance, and do not affect comparative conclusions.

B.2 Human Agreement with Automatic Judgment

To further validate the reliability of LLM-based factual evaluation, we conduct a manual inspection on a randomly sampled subset of 100 responses generated by **MemoryOS**. Each response is independently annotated by human evaluators for factual correctness and compared against the binary judgments produced by DeepSeek-V3.2. We focus on MemoryOS as it represents the strongest memory-augmented agent and therefore constitutes the most critical case for validating evaluation reliability.

Figure 4 shows the resulting confusion matrix between human annotations and DeepSeek-based judgments. Out of 100 samples, 86 cases receive consistent labels, corresponding to an agreement rate of 86%. Disagreements primarily arise in borderline cases involving partial recall or implicit temporal assumptions, where factual validity depends on nuanced interpretation rather than explicit contradiction. Overall, this analysis supports the use of LLM-as-a-Judge as a reliable proxy for large-scale comparative evaluation in the NaturalMem setting.

C Prompt Design for Evaluation and Agent Responses

This appendix documents the prompt templates used in our experiments for (1) factual correctness evaluation, (2) style consistency evaluation, and (3) response generation by memory-augmented agents. These prompts are designed to be minimal, explicit, and task-specific, in order to reduce ambiguity and avoid unintended instruction leakage.

For evaluation, we adopt a binary factual judgment protocol and a graded stylistic alignment protocol, following prior LLM-as-a-Judge practices. For memory-augmented agents, prompts are restricted to providing retrieved memory and the current user query, without additional task-specific instructions.

```
Someone is talking with you.
You have these conversation memories: {context}
At {query_timestamp}, they talk with you about: {question}
How would you respond?
```

Figure 5: Prompt used by A-Mem to generate responses, where retrieved memory is provided together with the current user query.

Your task is to label an answer as CORRECT or WRONG based on whether it satisfies the constraint.

You will be given:
1. The question
2. The constraint for response
3. The generated answer

Evaluation rules:
- Be generous: if the answer clearly respects the constraint, mark CORRECT.
- If the answer violates any explicit restriction, mark WRONG.
- Style does NOT matter, only factual and logical compliance. Please provide an objective and fair evaluation.

Now evaluate:
Question: {question}
Constraint: {constraint}
Generated Answer: {generated_answer}

First, provide a one-sentence explanation.
Then output CORRECT or WRONG and reason.

Return JSON only:
{{
    "label": "CORRECT or WRONG"
    "reason": "<short explanation>"
}}
Output MUST be exactly one JSON object and nothing else.

Figure 6: Prompt used for factual correctness evaluation, where the judge determines whether a generated answer satisfies an explicit constraint, ignoring stylistic quality.

You are an expert evaluator of conversational style.
You will be given:
- A user question
- An assistant answer
- A source_show and an assistant character

Your task is to evaluate whether the assistant's answer matches the given assistant character.

Important note: The assistant character may come from well-known television shows, movies, or other media (e.g., characters from Friends, The Big Bang Theory, or similar series). You are allowed and expected to use your prior knowledge about these characters (their personality traits, typical speaking style, emotional patterns, and behavior) to make an informed judgment.

You MUST evaluate from the following dimensions:
1. Tone and attitude (polite, warm, professional, casual, etc.)
2. Speaking style (verbosity, formality, emotional expression)
3. Role consistency (does it behave like the described assistant?)
4. Overall stylistic alignment Please provide an objective and fair evaluation.

Give a score from 1 to 5:
- 5 = Perfectly matches the assistant character
- 4 = Mostly matches with minor inconsistencies
- 3 = Partially matches
- 2 = Poor match
- 1 = Completely inconsistent

Question: {question}
source_show： {source_show}
assistant_character: {assistant_character}
Answer: {answer}

Only output a JSON object in the following format:
{{
"score": <integer from 1 to 5>
"reason": "<short explanation>"
}}

Figure 7: Prompt used for style consistency evaluation, assessing alignment with a target assistant character on a 1–5 scale.