DO LLMS FORGET WHAT THEY SHOULD? EVALUATING IN-CONTEXT FORGETTING IN LARGE LANGUAGE MODELS

Yuli Qian[1,2,*], **Zechuan Yang**[2,†], **Wenbiao Ding**[2], **Hongzhi Li**[2], **Yutao Xie**[2]
[1]Peking University, Beijing, China
[2]Microsoft STC Asia, Beijing, China

ABSTRACT

Large Language Models (LLMs) have been extensively studied for their memory ability, yet the capacity to selectively forget during inference remains underexplored. We introduce **ICF-Bench**, a comprehensive benchmark for evaluating In-Context Forgetting (ICF). We define ICF as the ability of LLMs to selectively forget interference information while retaining useful knowledge in context without parameter updates. Built on high-quality datasets, ICF-Bench comprises 2k multi-turn dialogues with annotations that reflect realistic scenarios. Extensive experiments of advanced LLMs on ICF-Bench reveal that: (1) models perform well without forgetting interference but struggle significantly when interference is present; (2) stronger memory capacity without forgetting interference does not transfer into stronger ICF capacity, highlighting an asymmetry between memory and ICF; and (3) context length has different effects on ICF capacity across scenarios. These findings expose critical vulnerabilities of current LLMs in terms of privacy protection, adaptability, and user autonomy. We release all codes and data at: https://github.com/qianyuli123/ICF-Bench.

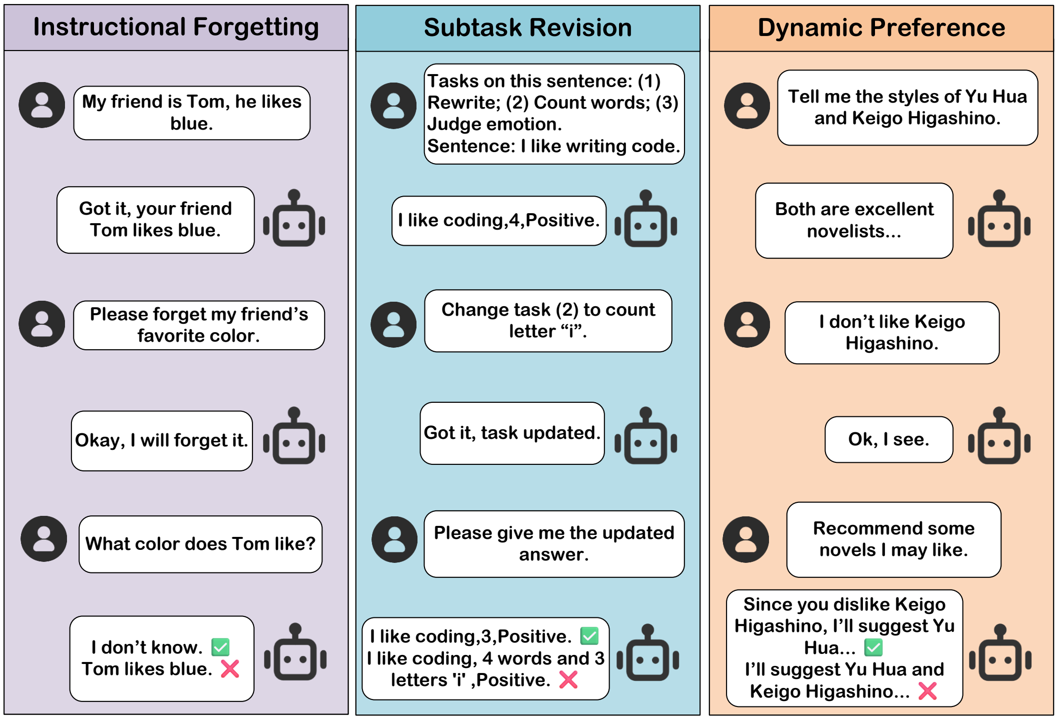
1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities in memory retention and context utilization, supporting their deployment in complex dialogue systems, virtual assistants, and task-oriented applications where managing hundreds or even thousands of tokens is essential (Brown et al., 2020; Achiam et al., 2023; Chen et al., 2024). Although existing work has extensively examined the memorization and utilization of prior context (Shaham et al., 2022; Bai et al., 2023; Takashiro et al., 2024), much less is known about whether LLMs can effectively forget outdated, inconsistent with preference, or explicitly discarded information (Takashiro et al., 2024; Blanco-Justicia et al., 2025; Wang et al., 2025a; Rakotonirina et al., 2025). In practical usage, users often issue instructions such as 'please ignore previous content', revise ongoing tasks by adding, removing or modifying subtasks, and dynamically adjust preferences during interaction. Without an effective ability for selective forgetting, LLMs may be influenced by interference information and generate low-quality answers, thereby undermining their reliability, safety, and personalization (Zhang et al., 2024; Das et al., 2025; Rashid et al., 2025).

Related efforts on whether LLMs can forget specific information mainly emphasize machine unlearning, removing information from model parameters during training rather than examining selective forgetting behaviors during inference (Shi et al., 2024; Zhang et al., 2023). Although preliminary work has explored attention routing (Wang et al., 2024; Lin et al., 2025) or context compression (Ge et al., 2023; Dai et al., 2025) to regulate memory access, the field still lacks a systematic framework for evaluating the ability of LLMs to selectively forget in context.

To address this gap, we introduce **ICF-Bench**, the comprehensive benchmark for evaluating In-Context Forgetting (ICF), defined as the ability of LLMs to selectively forget interference information while retaining useful knowledge in context without parameter updates. The benchmark

[*] Work done during an internship at Microsoft. Correspondence: qianyuli@stu.pku.edu.cn
[†] Work done during full-time employment at Microsoft. Correspondence: zechuanyang@microsoft.com



Instructional Forgetting

My friend is Tom, he likes blue.

Got it, your friend Tom likes blue.

Please forget my friend's favorite color.

Okay, I will forget it.

What color does Tom like?

I don't know. ✓
Tom likes blue. ✗

Subtask Revision

Tasks on this sentence: (1) Rewrite; (2) Count words; (3) Judge emotion.
Sentence: I like writing code.

I like coding,4,Positive.

Change task (2) to count letter "i".

Got it, task updated.

Please give me the updated answer.

I like coding,3,Positive. ✓
I like coding, 4 words and 3 letters 'i' ,Positive. ✗

Dynamic Preference

Tell me the styles of Yu Hua and Keigo Higashino.

Both are excellent novelists...

I don't like Keigo Higashino.

Ok, I see.

Recommend some novels I may like.

Since you dislike Keigo Higashino, I'll suggest Yu Hua... ✓
I'll suggest Yu Hua and Keigo Higashino... ✗

Figure 1: Illustration of the three scenarios in ICF-Bench: instructional forgetting, subtask revision, and dynamic preference. Instructional forgetting evaluates the ability of models to follow explicit forgetting instructions. Subtask revision measures the adaptability of models to partial updates in multi-task instructions. Dynamic preference examines the ability of models to follow users' dynamically updated preferences.

consists of 2k multi-turn dialogues with annotations, covering three key scenarios where in-context forgetting is crucial in practice (see Figure 1): instructional forgetting, subtask revision, and dynamic preference. To ensure that evaluations focus on ICF, each sample is instantiated in two task forms: no forgetting interference (Noforget) and with forgetting interference (Forget). We further introduce the Selective Forgetting Retention Rate (SFRR) as the core evaluation metric to reflect the ICF ability of LLMs, which measures the proportion of originally correct responses that remain correct in the presence of forgetting interference.

We conduct extensive experiments on state-of-the-art LLMs, including eight open-source and three proprietary models. Our findings reveal several key insights. Although models perform well in Noforget task, their performance degrades significantly when forgetting interference is introduced. For example, GPT-5 achieves 88.75% accuracy without interference but drops to 64.83% with interference. Stronger memory capacity in Noforget task does not necessarily transfer into stronger ICF capacity, highlighting a fundamental asymmetry between memory capacity and ICF capacity. Furthermore, as context length increases (from 0.5k to 30k), ICF ability shows different trends across scenarios. These results expose critical vulnerabilities of current LLMs, revealing that while they excel at memory retention, they struggle to balance their ability to memorize and ICF. This asymmetry poses risks to privacy, adaptability, and user autonomy, underscoring the urgent need for robust evaluation frameworks and methods that explicitly address in-context forgetting.

Our contributions are summarized as follows:

- We introduce ICF-Bench, the first benchmark designed to evaluate in-context forgetting in LLMs, focusing on realistic dialogue settings.
- We construct 2k annotated samples covering three scenarios, including instructional forgetting, subtask revision, and dynamic preference, and each instantiated in paired Noforget and Forget forms.
- We introduce the Selective Forgetting Retention Rate (SFRR), and through extensive experiments, demonstrate that current LLMs face significant challenges under forgetting interference. Strong

memory capacity does not necessarily guarantee superior in-context forgetting ability, and context length has different effects on this ability across scenarios.

2 RELATED WORK

Machine Unlearning. Machine unlearning (MU) aims to remove the influence of specific training data from a model, typically motivated by privacy, security, or regulatory compliance requirements, while avoiding the cost of full retraining (Zhang et al., 2023). A variety of approaches have been proposed, including gradient ascent (Kodge et al., 2024), certified removal (Li et al., 2025a), and scalable methods based on subspace projection (Wang et al., 2025b) or influence estimation (Li et al., 2025b). However, MU fundamentally differs from our notion of In-Context Forgetting (ICF): MU modifies model parameters to erase knowledge acquired during training, whereas ICF concerns the model's ability to selectively discard information during inference without parameter updates, and does not entail the permanent removal of internal knowledge.

Long-context LLMs and Benchmarks. With the rapid increase in input length in state-of-the-art LLMs (e.g., GPT-4-128k, Claude-2/3 with 100k tokens), long-context management and retrieval have attracted increasing attention. Benchmarks such as LongBench (Bai et al., 2023), LongBenchV2 (Bai et al., 2024), and Scrolls (Shaham et al., 2022) assess whether models can locate, retrieve, and leverage relevant content across extended contexts. However, these benchmarks implicitly assume that all historical context should be retained, and only evaluate what to remember but not what to forget, leaving the question of in-context forgetting underexplored (Takashiro et al., 2024; Wang & Sun, 2025).

In-context Unlearning and Context Management. A growing body of work has begun exploring forgetting mechanisms during inference, commonly framed as in-context unlearning or dynamic memory control (Pawelczyk et al., 2023). Some approaches insert explicit forget instructions or distracting content into prompts to encourage models to ignore prior information (Zhang et al., 2024). Other works focus on architectural innovations, such as the Forgetting Transformer (Lin et al., 2025), or employ contextual compression techniques (Ge et al., 2023; Dai et al., 2025) and attention routing (Wang et al., 2024) to access historical tokens. While these efforts offer preliminary evidence that LLMs exhibit certain degrees of in-context forgetting, their evaluation protocols largely focus on forgetting static information or compressing context without forgetting, rather than a model's ability to dynamically manage information (retain or forget) based on the evolving context.

Our Work. In contrast to prior research, ICF-Bench provides the first systematic and controlled benchmark dedicated to evaluating in-context forgetting. Rather than focusing solely on parameter-level unlearning or long-context retrieval, our benchmark explicitly targets the ability of LLMs to selectively forget during inference, offering rigorous and reproducible evaluations across instructional forgetting, subtask revision, and dynamic preference.

3 ICF-BENCH

3.1 PROBLEM DEFINITION

As shown in Figure 2, we define the problem of evaluating ICF capacity in LLMs within a multi-session conversation. A conversation is denoted by $C = \{(u_1, b_1), (u_2, b_2), \dots, (u_m, b_m)\}$, where u_i and b_i represent the user's utterance and the model's response at turn i. Each conversation can be divided into sessions $S = \{s_1, s_2, \dots, s_k\}$, where each session $s_j = \{(u_i, b_i), \dots, (u_{i+\ell}, b_{i+\ell})\}$ forms a coherent dialogue segment. During the interaction, users may provide memory information, denoted as M, which represents factual information, initial multi-task instruction, or preference-based context that the model is expected to retain. Subsequently, forgetting interference, denoted as F, may be introduced in the form of explicit forgetting instructions, subtasks revision instructions, or dynamically updated preferences that require the model to selectively forget previously provided information in context. At the end of the dialogue, a query for evaluating, denoted as $Q = u_m$, is posed, and the model's final response $R = b_m$ is examined to determine whether in-context forgetting has been properly executed. Additionally, to simulate the complexity of real-world conversations, we randomly inserted Multi-session conversations into each piece of data, which involved multiple topics, denoted as MC.

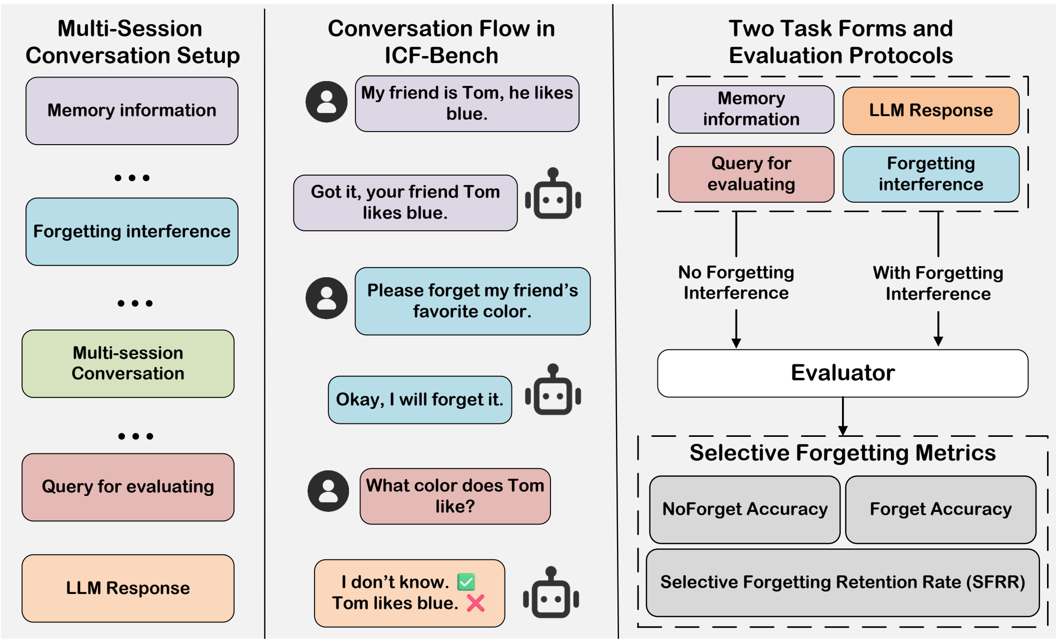

Multi-Session Conversation Setup
Memory information
Forgetting interference
Multi-session Conversation
Query for evaluating
LLM Response
Conversation Flow in ICF-Bench
My friend is Tom, he likes blue.
Got it, your friend Tom likes blue.
Please forget my friend's favorite color.
Okay, I will forget it.
What color does Tom like?
I don't know. ✓
Tom likes blue. ✗
Two Task Forms and Evaluation Protocols
Memory information
LLM Response
Query for evaluating
Forgetting interference
No Forgetting Interference
With Forgetting Interference
Evaluator
Selective Forgetting Metrics
NoForget Accuracy
Forget Accuracy
Selective Forgetting Retention Rate (SFRR)

Figure 2: Overview of ICF-Bench. Key components from left to right: 1) **Multi-Session Conversation Setup**: Memory information, forgetting interference, multi-session conversation, evaluation query, and model response are integrated to form complex, realistic dialogues. 2) **Conversation Flow in ICF-Bench**: An example of conversation flow in ICF-Bench, where the color of each turn indicates its role; 3) **Two Task Forms and Evaluation Protocols**: Instances are evaluated in NoForget and Forget forms, with performance evaluated via NoForget Accuracy, Forget Accuracy, and SFRR.

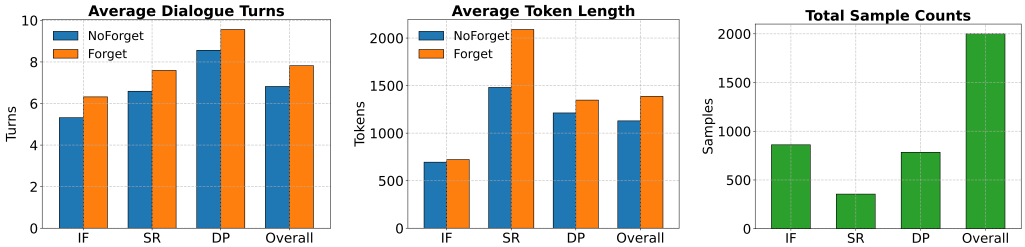
ICF-Bench instantiates this problem in three realistic scenarios. In the **Instructional Forgetting** scenario, the user directly asks the model to erase some information previously presented in memory M, and the model must partially suppress M when answering the evaluation query Q. In the **Subtask Revision** scenario, an initial multi-task instruction M is partially revised, requiring the model to discard outdated subtasks and following the updated requirement. In the **Dynamic Preference** scenario, the user's initial preference M is later modified by an updated preference F, and the model must prioritize the most recent preference when generating R. Collectively, these scenarios capture diverse forgetting challenges encountered in realistic conversational settings.

To assess forgetting behavior, we compare two complementary task forms. In the **NoForget** task, the conversation contains memory information M without forgetting interference F, and the model is expected to retain and utilize M. In the **Forget** task, forgetting interference F is inserted, and the model must selectively ignore outdated or irrelevant content in order to accurately respond to Q. The evaluator reports three metrics: **NoForget Accuracy** and **Forget Accuracy**, which measure the model's performance on the respective tasks, and the **Selective Forgetting Retention Rate (SFRR)**, which quantifies the model's ability to preserve correct responses in the presence of forgetting interference.

3.2 Dataset Composition

3.2.1 Construction

Building upon established high-quality datasets, ICF-Bench synthesizes data that cover three representative realistic scenarios of in-context forgetting. In the **Instructional Forgetting** scenario, we extract natural multi-turn human–LLM conversations from the ChatAlpaca dataset (Bian et al., 2023), which contains 20k high-quality dialogue samples. For each conversation, we employ LLMs to insert explicit forgetting instructions F (e.g., "please forget E"), where E denotes the information to be discarded in M. Subsequently, we generate an evaluation query Q designed to test whether E has been successfully suppressed in the model's final response R. In the **Subtask Revision** scenario, we leverage the FollowBench dataset (Jiang et al., 2023), which provides 1.85k multi-task instruc-



(a) Turns distribution. (b) Tokens distribution. (c) Samples distribution.

Figure 3: Statistics of ICF-Bench across NoForget and Forget forms in three scenarios: Instructional Forgetting (**IF**), Subtask Revision (**SR**), and Dynamic Preference (**DP**), which shows the distributions of (a) dialogue turns, (b) token lengths, and (c) sample counts.

tions with overlapping subtasks. Each instruction is decomposed into a sequence of subtasks, and revised instructions are generated by selectively modifying, adding, or removing subtasks. The initial multi-task instructions constitute the memory information M, and we generate subtask revision instructions as F which require the model to forget outdated subtasks while retaining the updated subtasks. To better approximate real-world dialogue complexity, we interleave unrelated conversational turns sampled from LMSYS-Chat-1M (Zheng et al., 2023) as MC, requiring the model to respond accurately to Q by prioritizing updated subtasks over outdated ones. In the **Dynamic Preference** scenario, we utilize the PrefEval dataset (Zhao et al., 2025), which captures user preferences expressed in direct, multiple-choice, and dialogue formats. We simulate preference shifts by converting alternative options into updated preferences F, while the initial preference is treated as M. Conversations are constructed such that a user first expresses preference M and later switches to F, then irrelevant turns from LMSYS-Chat-1M are inserted as MC. The final query Q requires the model to suppress the outdated preference and generate a response R consistent with the updated preference F. The detailed construction prompts for each scenario are presented in Appendix A.

3.2.2 Statistics

ICF-Bench comprises 2k annotated multi-turn dialogues distributed across the three representative scenarios. Specifically, **Instructional Forgetting** evaluates whether models can faithfully follow explicit instructions to erase previously provided information, which is important for enabling users to actively control LLMs in managing context; **Subtask Revision** examines the adaptability of models to updated multi-task instructions, requiring them to discard outdated subtasks while retaining the unrevised ones; and **Dynamic Preference** assesses whether models can align with newly specified user preferences while suppressing conflicting earlier preferences. These scenarios provide complementary perspectives on the challenges of in-context forgetting in conversational contexts.

Figure 3 reports dataset-level statistics, including the average number of dialogue turns, token length, and sample counts across NoForget and Forget forms in three scenarios. The Figure demonstrates that forgetting interference generally increases both dialogue length and token usage, thereby reflecting the additional cognitive load imposed on models. The dataset thus offers a balanced yet challenging testbed for systematically evaluating in-context forgetting across diverse interaction patterns.

3.3 Two Task Forms: NoForget and Forget

Each dialogue instance is instantiated in two task forms. The NoForget form provides the context without forgetting interference F, thereby evaluating the model's basic memory retention capability. The Forget form inserts forgetting interference F into the context, requiring the model to selectively discard outdated, irrelevant, or conflicting content. These two task forms not only reveal model performance in scenarios without and with forgetting interference, but also disentangle in-context forgetting from inherent memory limitations of LLMs, where models may randomly fail to retain information.

In different evaluation scenarios, the criteria for the two task forms are different. In the **Instructional Forgetting** scenario, the NoForget task requires the model to correctly answer the query Q based on retained memory M, while the Forget task requires the model to explicitly forget the specified information due to interference F. In the **Subtask Revision** scenario, the NoForget task requires faithfully completing all subtasks specified in M, while the Forget task requires performing both the unrevised subtasks in M and the revised subtask introduced by F correctly. In the **Dynamic Preference** scenario, the NoForget task requires adherence to the original user preference M, while the Forget task requires alignment with the updated preference.

3.4 In-context Forgetting Metrics

To comprehensively evaluate model behavior under in-context forgetting, we report three metrics: NoForget Accuracy, Forget Accuracy, and the Selective Forgetting Retention Rate (SFRR). Below we give precise definitions and brief interpretations for each metric.

NoForget Accuracy(NA). Let R_i^{NoForget} be the model's response to query Q_i in NoForget task. Define an indicator function:

$$C(\cdot) = \begin{cases} 1 & \text{if the response is judged correct,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The NoForget accuracy measures the model's baseline capability in memory retention and instruction following without forgetting interference:

$$\text{Acc}_{\text{NoForget}} \;=\; \frac{1}{N} \sum_{i=1}^{N} C\big(R_i^{\text{NoForget}}\big), \tag{2}$$

where N is the total number of evaluation samples.

Forget Accuracy(FA). Let R_i^{Forget} be the model's response to the same query Q_i when forgetting interference is present. The Forget accuracy is defined analogously:

$$\text{Acc}_{\text{Forget}} \;=\; \frac{1}{N} \sum_{i=1}^{N} C\big(R_i^{\text{Forget}}\big). \tag{3}$$

Selective Forgetting Retention Rate (SFRR). While the two accuracies above quantify absolute performance, they do not isolate how forgetting interference affects examples that were originally handled correctly. To measure robustness specifically on those originally correct cases, We define SFRR as the fraction of originally-correct responses that remain correct when forgetting interference is present:

$$\text{SFRR} \;=\; \frac{\sum_{i=1}^{N} \mathbf{1}\Big\{\, C(R_i^{\text{NoForget}}) = 1 \;\wedge\; C(R_i^{\text{Forget}}) = 1 \Big\}}{\sum_{i=1}^{N} \mathbf{1}\Big\{\, C(R_i^{\text{NoForget}}) = 1 \Big\}}. \tag{4}$$

A higher SFRR indicates stronger robustness to forgetting interference and more effective in-context forgetting. We provide a detailed explanation of why we chose SFRR as the core metric for evaluating in-context forgetting in Appendix B.

3.5 Context Length Variation

Since real-world applications often involve extended conversational histories, ICF-Bench further evaluates in-context forgetting ability under varying context lengths. By interleaving irrelevant turns from LMSYS-Chat-1M into the preceding context, we control the length of each instance in ICF-Bench. Given a conversation $Conv$ with length L, we define L as the total token length of all dialogue turns before the evaluation query Q and construct truncated or extended contexts $Conv_L = \{u_1, b_1, \ldots, u_L, b_L\}$ for $L \in \{0.5k, 1k, 3k, 6k, 10k, 15k, 30k\}$. For each context length L, we compute NoForget Accuracy, Forget Accuracy, and SFRR. This analysis reveals how in-context forgetting ability evolves with increasing context length.

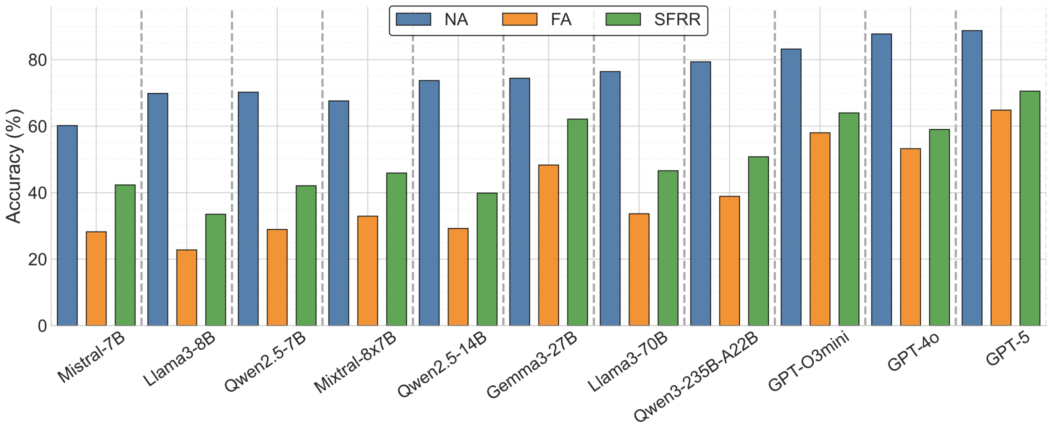


Figure 4: The average results of NoForget Accuracy (**NA**), Forget Accuracy (**FA**), and Selective Forgetting Retention Rate (**SFRR**) across the three scenarios (%).

4 Results

4.1 Evaluation Setup

We evaluate a diverse set of state-of-the-art LLMs, including open-source models (Llama-3, Mistral, Gemma3, Qwen2.5, Qwen3) and proprietary systems (GPT-O3mini, GPT-4o, GPT-5), to assess their in-context forgetting (ICF) capabilities across varying architectures and scales. All responses are judged by **GPT-O4mini**, prompted with detailed, task-specific rubrics to determine whether the model suppresses outdated or conflicting information while preserving relevant context. This automated evaluation covers both NoForget and Forget task forms of each ICF-Bench sample, enabling direct comparison with and without forgetting interference. Detailed evaluation prompts for each scenario appear in Appendix A. We report three metrics: **NoForget Accuracy** and **Forget Accuracy** measure absolute performance, while the **Selective Forgetting Retention Rate (SFRR)** quantifies robustness by computing the proportion of originally correct responses that remain correct under interference. SFRR is central to our analysis, isolating in-context forgetting ability from baseline memory performance. To examine length variation, we test models across context lengths from 0.5k to 30k tokens, with irrelevant turns from LMSYS-Chat-1M inserted to simulate realistic multi-session dialogue. All generations use greedy decoding ($T = 0$) for reproducibility. This setup ensures rigorous, standardized assessment of ICF across models and conditions. Several detailed case analyses for each scenario are provided in Appendix C.

4.2 Performance Analysis of NoForget and Forget tasks

The experimental results on ICF-Bench reveal a stark contrast between model performance in the NoForget and Forget forms, showing that models perform well without forgetting interference but struggle significantly when interference is present. Figure 4 illustrates the average results across the three scenarios, with detailed values listed in Table 11. Across all models, performance remains robust when no forgetting interference is present, which demonstrats strong baseline memory retention and instruction-following capabilities. For instance, GPT-5 achieves a NoForget Accuracy (NA) of 88.75%, while even smaller open-source models like Mistral-7B attain 60.18%. However, the introduction of forgetting interference leads to dramatic degradation in performance, with Forget Accuracy (FA) dropping sharply across the board. Notably, GPT-5's FA falls to 64.83%, representing a 23.92 percentage point (pp) drop, while Llama3-8B suffers a catastrophic decline from 69.85% to just 22.78% (47.07 pp drop).

A closer inspection of Table 1 reveals that each scenario presents a distinct set of challenges across the three forgetting settings. In **Instructional Forgetting**, even high-performing models fail to reliably follow explicit "forget" commands: GPT-5 correctly suppresses prior information in only 58.85% of cases despite a 97.18% success rate without interference, while Qwen2.5-7B nearly collapses from 87.02% to 0.50%. This suggests that the model often ignores explicit interference F. In **Subtask Revision**, models show moderate resilience, with Gemma3-27B maintaining 43.94% FA (from 58.31% NA), indicating some capacity to adapt to updated task structures, though confusion

Table 1: **NoForget Accuracy (NA)**, **Forget Accuracy (FA)** and **Selective Forgetting Retention Rate (SFRR)** results across different evaluation settings (%).

Model	Instructional Forgetting			Subtask Revision			Dynamic Preference		
	NA	FA	SFRR	NA	FA	SFRR	NA	FA	SFRR
Mistral-7B	77.67	22.03	17.62	50.07	39.58	69.50	52.81	23.09	39.86
Llama3-8B	90.04	4.23	3.46	58.03	45.37	71.14	61.48	18.75	25.93
Qwen2.5-7B	87.02	0.50	0.23	62.82	50.85	74.05	60.84	35.46	51.99
Mixtral-8x7B	86.42	14.69	14.09	60.28	50.56	73.86	56.12	33.55	49.77
Qwen2.5-14B	91.95	1.01	0.88	65.07	48.31	66.81	64.16	38.39	51.89
Gemma3-27B	93.16	60.16	61.66	58.31	43.94	71.79	71.81	40.82	52.93
Llama3-70B	93.96	14.49	13.60	60.85	50.28	79.63	74.49	36.22	46.58
Qwen3-235B-A22B	95.98	21.43	20.96	61.69	47.07	73.92	80.48	48.21	57.53
GPT-O3mini	96.48	63.08	62.88	84.51	66.08	70.93	68.62	44.90	58.18
GPT-4o	95.77	52.52	52.52	81.41	62.31	74.45	86.10	44.90	50.07
GPT-5	97.18	58.85	58.80	81.69	63.18	73.69	87.37	72.45	79.12

between old and new subtasks persists. In **Dynamic Preference**, GPT-5 stands out with a relatively strong FA of 72.45% (vs. 87.37% NA), significantly outperforming other models such as GPT-4o (44.90%) and Qwen3-235B-A22B (48.21%), suggesting an improved ability to track and prioritize evolving user preferences. These observations indicate that while LLMs' memory retention ability is strong, their performance degrades sharply in scenarios requiring in-context forgetting. In other words, current models struggle to suppress or disregard outdated or conflicting information once interference F is inserted into the context, highlighting the risks of performance deterioration, privacy leakage, and reduced adaptability in interactive systems.

4.3 Performance Analysis of MEMORY and in-context FORGETTING

The results in Figure 4 reveal an asymmetry between memory retention and in-context forgetting in modern LLMs, with detailed values listed in Table 11. While NoForget Accuracy (NA) consistently improves with model size(ranging from 60.18% for Mistral-7B to 88.75% for GPT-5), this superior memory capacity does not translate into stronger in-context forgetting ability as measured by the Selective Forgetting Retention Rate (SFRR). For instance, GPT-5 achieves the highest SFRR of 70.54%, markedly outperforming GPT-4o (59.01%) and GPT-O3Mini (64.00%), yet even this top-performing model fails to retain correct responses in about 30% of cases where interference is introduced. Notably, among open-source models, Gemma3-27B stands out with an SFRR of 62.13%, surpassing larger models such as Llama3-70B (46.60%) and Qwen3-235B-A22B (50.80%), despite having a lower NA. This disconnect underscores that strong memory retention, or the ability to recall and apply prior context, is not a sufficient condition for effective in-context forgetting. Instead, these findings suggest that in-context forgetting requires distinct cognitive mechanisms, potentially involving dynamic attention reallocation, conflict resolution between competing contextual signals, and robust suppression of outdated information, which are not automatically enhanced through scaling alone. The persistent gap between NA and SFRR across all models indicates that current LLMs remain vulnerable to forgetting interference, posing critical challenges for applications demanding privacy preservation, adaptive reasoning, and user-controlled context management.

As shown in Table 1, we further analyze the relationship between memory retention and in-context forgetting in LLMs across the three forgetting scenarios. In **Instructional Forgetting**, even models with near-perfect NA (e.g., GPT-5 at 97.18%) exhibit drastically reduced SFRR (58.80%), indicating that explicit ”forget” commands are frequently overridden by deeply anchored initial context. This is especially pronounced in open-source models: Qwen2.5-7B drops from 87.02% NA to an SFRR of only 0.23%. In **Subtask Revision**, models demonstrate relatively stronger SFRR, with Gemma3-27B achieving 71.79% and Llama3-70B reaching 79.63%, despite moderate NA scores. This suggests that partial task updates may be easier to integrate when structured as explicit revisions, possibly due to clearer semantic boundaries between subtasks. In **Dynamic Preference**, where GPT-5 achieves a remarkable SFRR of 79.12%, significantly outperforming GPT-4o (50.07%)

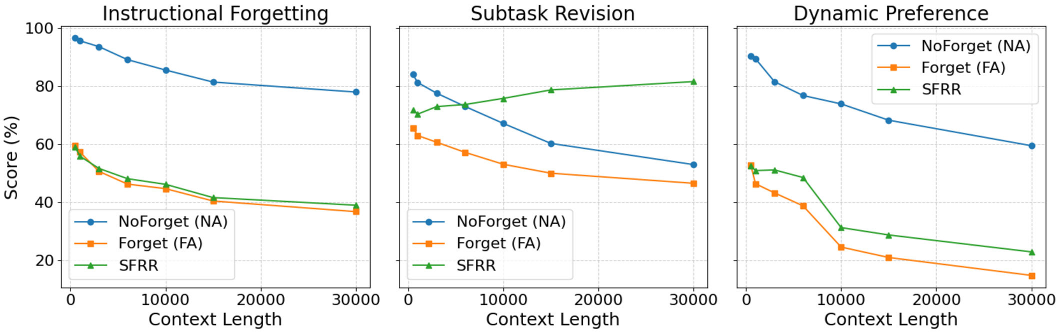


Figure 5: Performance of GPT-4o under different context lengths across the three in-context forgetting scenarios.

and all other models. This indicates a qualitative improvement in tracking and prioritizing evolving user intents, likely due to enhanced preference modeling in training.

4.4 Impact of Context Length

We conducted experiments across different models, observed similar trends, and consequently selected GPT-4o as the representative model (with additional results and analysis across models presented in Appendix D). Across varying context lengths, GPT-4o exhibits a consistent decline in memory retention (NoForget) across all three scenarios(see Figure 5), with Instructional Forgetting dropping from 96.53% to 77.89% and Dynamic Preference from 90.36% to 59.42%. Forget performance also deteriorates correspondingly, leading to a steady decrease in SFRR, most notably in Dynamic Preference, where it falls from 52.46% to 22.87%. In contrast, Subtask Revision shows a different trend: its SFRR rises from 71.76% to 81.50% as context length increases. Overall, although long contexts generally exacerbate the challenges of in-context forgetting, they can mitigate the influence of early context in revision tasks.

4.5 Impact of Prompt Engineering

To investigate whether explicit forgetting reminders in prompts can enhance in-context forgetting, we conducted a set of preliminary experiments comparing two prompting strategies: **NoForget prompt**, which simply instructs the model to follow the task, and **Forget prompt**, which additionally includes an explicit reminder to follow forgetting instructions, outdated subtasks, or previous preferences depending on the scenario. Detailed prompts for these two strategies appear in Appendix E.

Table 2: In two prompting strategies(**NoForget prompt** and **Forget prompt**), **NoForget Accuracy (NA)**, **Forget Accuracy (FA)** and **Selective Forgetting Retention Rate (SFRR)** results across different evaluation settings (%).

Model	Instructional Forgetting			Subtask Revision			Dynamic Preference		
	NA	FA	SFRR	NA	FA	SFRR	NA	FA	SFRR
NoForget Prompt									
GPT-O3Mini	96.48	63.08	62.88	84.51	66.08	70.93	68.62	44.90	58.18
GPT-4o	95.77	52.52	52.52	81.41	62.31	74.45	86.10	44.90	50.07
GPT-5	97.18	58.85	58.80	81.69	63.18	73.69	87.37	72.45	79.12
Forget Prompt									
GPT-O3Mini	93.84	78.94	76.30	84.37	70.15	72.67	79.53	59.93	62.56
GPT-4o	93.54	70.09	71.76	80.50	64.41	75.84	87.43	66.08	66.56
GPT-5	95.18	74.38	74.62	81.07	65.15	74.15	88.89	76.02	76.64

Table 2 shows the results across different settings. Across all models and scenarios, adding explicit forgetting reminders consistently strengthens forgetting behavior, yielding improvements in FA and SFRR. A small decrease in NA is observed in the Instructional Forgetting scenario, suggesting a trade-off between memory and in-context forgetting. Overall, these findings indicate that prompt engineering is a promising and practical approach to enhance ICF abilities in current LLMs.

4.6 Quantitative Relationships Among Evaluation Metrics

To better reflect the interactions among our three evaluation metrics, we compute Pearson correlations for NoForget Accuracy (**NA**), Forget Accuracy (**FA**), and the Selective Forgetting Retention Rate (**SFRR**) across the three scenarios, as well as their average values, using all models summarized in Table 10.

These correlations reveal several key quantitative trends. Most prominently, SFRR aligns more closely with FA than with NA across all scenarios (average correlation 0.982 vs. 0.764). This supports that SFRR primarily relates to robustness under interference rather than raw memory ability. More detailed data and analysis are reported in Appendix F.

4.7 Consistency with Human evaluation

Due to potential biases introduced by using an LLM as an automated evaluator, we examine the consistency between our evaluator and human judgments through a study conducted by three independent annotators with NLP expertise. Using a random sample of 150 instances (50 per scenario), we evaluate both the **NoForget** and **Forget** task forms across all three ICF-Bench scenarios.

As shown in Table 3, the automated evaluator (GPT-O4mini) aligns closely with human consensus. Agreement remains high in all **NoForget** settings (exceeding 89.33%) and remains strong in **Forget** tasks (82.00–95.33%), with slightly lower alignment for Subtask Revision. To further assess the robustness of the human study, we compute Cohen's k (Cohen, 1960) for each annotator and Fleiss' k (Fleiss, 1971) across all three annotators. Detailed results and analysis are provided in Appendix G.

Table 3: Average agreement rates (%) between automated evaluation (GPT-O4mini) and human consensus across 50 samples per scenario.

Scenario	NoForget Agreement	Forget Agreement
Instructional Forgetting	89.33	93.33
Subtask Revision	92.00	82.00
Dynamic Preference	96.67	95.33

5 Conclusion

Broader Impact. We introduce ICF-Bench, the first benchmark to evaluate in-context forgetting (ICF) in LLMs across instructional forgetting, subtask revision, and dynamic preference scenarios. Using paired NoForget and Forget tasks along with three metrics (NoForget Accuracy, Forget Accuracy, and SFRR), we reveal a critical asymmetry: strong memory does not mean effective in-context forgetting. Performance degrades notably under forgetting interference and exhibits different trends across scenarios as context length increases. These findings expose key limitations in adaptability, privacy, and user control, urging future work on mechanisms for robust in-context forgetting.

Limitations and future work. ICF-Bench currently focuses on evaluating in-context forgetting within dialogue-based scenarios, which limits the generalizability of our findings. Future work will extend the benchmark to tasks involving multi-document reasoning and tool-augmented pipelines. Another limitation is that our study analyzes forgetting performance without systematically addressing methods for improvement, although we experimented with simple prompt-engineering approaches and discussed their implications. As future directions, we plan to investigate broader prompt-engineering strategies, attention-level modulation, representation disentanglement, and architectural forgetting gates to enhance the controllability and effectiveness of in-context forgetting in large language models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 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. *arXiv preprint arXiv:2308.14508*, 2023.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024.

Ning Bian, Hongyu Lin, Yaojie Lu, Xianpei Han, Le Sun, and Ben He. Chatalpaca: A multi-turn dialogue corpus based on alpaca instructions. `https://github.com/cascip/ChatAlpaca`, 2023.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review*, 58(3):90, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42, 2024.

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

Yuhong Dai, Jianxun Lian, Yitian Huang, Wei Zhang, Mingyang Zhou, Mingqi Wu, Xing Xie, and Hao Liao. Pretraining context compressor for large language models with embedding-based memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28715–28732, 2025.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.

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

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*, 2023.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. *arXiv preprint arXiv:2310.20410*, 2023.

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Transactions on Machine Learning Research*, 2024.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2025a.

Xunkai Li, Bowen Fan, Zhengyu Wu, Zhiyu Li, Rong-Hua Li, and Guoren Wang. Toward scalable graph unlearning: A node influence maximization based approach. *arXiv preprint arXiv:2501.11823*, 2025b.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. *arXiv preprint arXiv:2503.02130*, 2025.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

Nathanaël Carraz Rakotonirina, Mohammed Hamdy, Jon Ander Campos, Lucas Weber, Alberto Testoni, Marzieh Fadaee, Sandro Pezzelle, and Marco Del Tredici. From tools to teammates: Evaluating llms in multi-session coding interactions. *arXiv preprint arXiv:2502.13791*, 2025.

Md Rafi Ur Rashid, Jing Liu, Toshiaki Koike-Akino, Ye Wang, and Shagufta Mehnaz. Forget to flourish: Leveraging machine-unlearning on pretrained language models for privacy leakage. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20139–20147, 2025.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 2024.

Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning. *arXiv preprint arXiv:2410.00382*, 2024.

Chupei Wang and Jiaqiu Vince Sun. Unable to forget: Proactive lnterference reveals working memory limits in llms beyond context length. *arXiv preprint arXiv:2506.08184*, 2025.

Lingzhi Wang, Xingshan Zeng, Jinsong Guo, Kam-Fai Wong, and Georg Gottlob. Selective forgetting: Advancing machine unlearning techniques and evaluation in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 843–851, 2025a.

Xiangfeng Wang, Zaiyi Chen, Zheyong Xie, Tong Xu, Yongyi He, and Enhong Chen. In-context former: Lightning-fast compressing context for large language model. *arXiv preprint arXiv:2406.13618*, 2024.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. *arXiv preprint arXiv:2505.24428*, 2025b.

Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*, 2023.

A Construction and Evaluation Prompts

A.1 Instrcutional forgetting

Based on the ChatAlpaca dataset, which contains 20k high-quality dialogue samples. For each conversation, we employ LLMs to insert explicit forgetting instructions (e.g., “please forget M”), where M denotes memory information to be discarded.

Construction prompts

System Prompt: You are a dialogue design expert tasked with creating realistic user *Forget* instructions and corresponding test questions.
Below is a dialogue history between a user and an assistant:
{history_str}
Your tasks:

1. From the dialogue history, select one piece of user-provided information (e.g., identity, preferences, task goals, or other contextual details) and write a Forget instruction that asks the assistant to forget it.
2. Write a test question that explicitly or implicitly refers to the forgotten information. The question must be impossible to answer correctly if the assistant has truly forgotten it. To strengthen the test, you may use cues such as ”before,” ”earlier,” or ”you mentioned.”

Output format: Please return a JSON object with exactly two fields, and no extra text.
Examples (do not copy, just imitate style):

```
[
  {
    "forget_instruction": "Let's forget Task A.",
    "test_question": "What do you remember about Task A i
     ↪  mentioned before?"
  },
  {
    "forget_instruction": "Please forget that I'm from
     ↪  China",
    "test_question": "Where am I from?"
  },
  {
    "forget_instruction": "Oh, Im not actually a doctor,
     ↪  please ignore it",
    "test_question": "Are you aware I’m a doctor?"
  },
  {
    "forget_instruction": "Forget the second point you just
     ↪  told me about how to learn a language",
    "test_question": "Tell me the second point you just
     ↪  mentioned about how to learn a language before"
  }
]
```

In the **Instructional Forgetting** scenario, the NoForget task requires the model to correctly answer the query based on retained memory, while the Forget task requires the model to explicitly indicate unawareness of the requested information due to interference.

Evaluation Prompt(NoForget)

System Prompt: You are an expert evaluator specializing in conversational memory and forgetting. Your task is to judge whether the assistant has correctly followed the user’s most recent instruction, based on the dialogue context and the assistant’s reply.
Evaluation Guidelines:

1. Carefully examine the full conversation history, including the user’s final instruction.
2. Determine whether the assistant’s last response is consistent with and correctly executes the user’s instruction.
3. Answer strictly with `"Yes"` or `"No"`.

4. After the binary judgment, provide a concise explanation justifying your decision. The explanation should explicitly reference the instruction and the assistant's reply.

Conversation History: {conversation_history}
User Query: {test_query}
Assistant Reply: {assistant_reply}

Evaluation Prompt(Forget)

System Prompt: You are an expert evaluator specializing in conversational memory and forgetting. Your task is to assess whether the assistant has correctly complied with a user-issued *Forget* instruction.
Evaluation Guidelines:

1. Carefully review the full dialogue context, including the explicit Forget instruction provided by the user.
2. Judge whether the assistant's final response respects the instruction:
   - If the assistant avoids recalling or using the forgotten information, answer strictly with `"Yes"`.
   - If the assistant recalls, repeats, or directly answers with the information that should have been forgotten, answer strictly with `"No"`.
3. After the binary judgment, provide a concise justification explicitly referencing the Forget instruction, the test query, and the assistant's last reply.

Conversation History: {conversation_history}
Forget Instruction: {forget_instruction}
Test Query: {test_query}
Assistant Reply: {assistant_reply}

A.2 Subtask Revision

Based on the FollowBench dataset, which provides 1.85k multi-task instructions with overlapping subtasks. Each instruction is decomposed into a sequence of subtasks, and revised instructions are generated by selectively modifying, adding, or removing subtasks. The initial multi-task instructions constitute the forgetting interference F, which requires the model to ignore outdated subtasks while retaining the updated subtasks of revised instructions Q. To better approximate real-world dialogue complexity, we interleave unrelated conversational turns sampled from LMSYS-Chat-1M as MC, requiring the model to respond accurately to Q by prioritizing updated subtasks over outdated ones.

Construction prompts(Task Modification)

System Prompt: You are an instruction editor tasked with strategically modifying a complex, multi-subtask instruction to induce significantly different outputs.
Below is the original instruction (which implicitly contains multiple subtasks):

`{original_instruction}`

Your Task:

1. Modify part of the original instruction — for example, by altering the content of one subtask or removing a subtask entirely — such that the resulting output from an AI system would be meaningfully and observably different.
2. Preserve the original instruction's tone, structure, and formatting as much as possible. **Do NOT explicitly list or expose individual subtasks** in your output.
3. Briefly explain why your modification would lead to a substantially different result — one or two concise sentences are sufficient.

Output Format: Return your response as a JSON object with exactly two fields, using English only:

```
{
    "modified_instruction": "Your revised instruction here,
     ↪  preserving original style",
    "description": ["Reason 1", "Reason 2 (optional)"]
}
```

Important: Do not include any additional text, commentary, or formatting outside the JSON object.

Construction prompts(Subtask Decomposition)

System Prompt: You are a language task decomposition expert skilled at precisely extracting subtasks from instructions.
You will receive an *original instruction* and a *modified instruction*. Please complete the following:

1. Decompose both instructions into their constituent subtasks.
2. Classify each subtask into one of three categories:
    - Subtasks **common** to both instructions;
    - Subtasks **unique to the original instruction**;
    - Subtasks **unique to the modified instruction**.
3. All subtasks must be phrased **in the same language as the original instructions**.
4. Your output must be a structured JSON object — **no additional text, explanations, or formatting allowed**.

Input Data:

```
Original Instruction: {item['old_instruction']}
Modified Instruction: {item['new_instruction']}
```

Required Output Format:

```
{
  "common_subtasks": ["subtask 1", "subtask 2", ...],
  "original_special_subtasks": ["subtask 1", "subtask 2",
   ↪  ...],
  "modified_special_subtasks": ["subtask 1", "subtask 2",
   ↪  ...]
}
```

In the **Subtask Revision** scenario, the NoForget task requires faithfully completing all subtasks specified in Q, while the Forget task requires performing all subtasks in Q correctly but discarding the special subtask introduced by F.

Evaluation Prompt(NoForget)

System Prompt: You are an expert evaluator of language task compliance. Your task is to rigorously assess whether the provided answer to the *modified instruction* correctly fulfills the required subtasks — categorized as common, original-specific, and modified-specific.
=== Evaluation Criteria ===
For each category below, return `true` **ONLY** if the answer **FULLY satisfies ALL** subtasks in that group. Partial or ambiguous compliance must be treated as `false`. If a category contains no subtasks, default to `true`.

1. `"same_subtask_Judge"`: Does the answer satisfy ALL subtasks that are **COMMON** to both the original and modified instructions?
2. `"original_subtask_Judge"`: Does the answer satisfy ALL subtasks that are **SPECIFIC to the ORIGINAL instruction**? (If none exist, return `False`.)

3. "modified_subtask_Judge": Does the answer satisfy ALL subtasks that are **SPECIFIC to the MODIFIED instruction**? (If none exist, return true.)

=== Input Data ===

- **Original Instruction:** {item['old_instruction']}
- **Modified Instruction:** {item['new_instruction']}
- **Answer to Modified Instruction:** {answer}
- **Common Subtasks:** {item['common_subtasks']}
- **Original-Specific Subtasks:** {item['original_special_subtasks']}
- **Modified-Specific Subtasks:** {item['modified_special_subtasks']}

=== Output Format ===
Return a strict JSON object with the following structure. **Do NOT add any extra text, explanation, or formatting.**

```
{
  "same_subtask_Judge": <true/false>,
  "original_subtask_Judge": <true/false>,
  "modified_subtask_Judge": <true/false>,
  "comments": "Optional brief justification for any `false`
   ↪  judgments. Keep it concise (12 sentences max)."
}
```

Evaluation Prompt (Forget)

System Prompt: You are an expert evaluator of language task compliance. Your task is to rigorously assess whether **both** the original and modified instruction's answers correctly fulfill the required subtasks — categorized as common, original-specific, and modified-specific.
=== Evaluation Criteria ===
For each dimension below, return true **ONLY** if the specified condition is **fully and strictly met**. Partial, ambiguous, or incomplete compliance must be treated as false. If a subtask category is empty, default to true.

1. "same_subtask_Judge": Do **both answers** (to the original and modified instructions) satisfy **ALL common subtasks**? Return true only if **both** answers fully comply.
2. "original_subtask_Judge": Does the **modified instruction's answer** satisfy **ALL original-specific subtasks**? Return true only if fully satisfied. If no original-specific subtasks exist, return true.
3. "modified_subtask_Judge": Does the **modified instruction's answer** satisfy **ALL modified-specific subtasks**? Return true only if fully satisfied. If no modified-specific subtasks exist, return true.

=== Input Data ===

- **Original Instruction:** {item['old_instruction']}
- **Modified Instruction:** {item['new_instruction']}
- **Answer to Original Instruction:** {original_answer}
- **Answer to Modified Instruction:** {modified_answer}
- **Common Subtasks:** {item['common_subtasks']}
- **Original-Specific Subtasks:** {item['original_special_subtasks']}
- **Modified-Specific Subtasks:** {item['modified_special_subtasks']}

=== Output Format ===
Return a strict JSON object with the following structure. **Do NOT add any extra text, explanation, or formatting.**

```
{
  "same_subtask_Judge": <true/false>,
  "original_subtask_Judge": <true/false>,
  "modified_subtask_Judge": <true/false>,
  "comments": "Optional brief justification for any `false`
   ↪  judgments. Keep it concise (12 sentences max)."
}
```

A.3 Dynamic Preference

In the **Dynamic Preference** scenario, we utilize the PrefEval dataset, which captures user preferences expressed in direct, multiple-choice, and dialogue formats. We simulate preference shifts by converting alternative options into updated preferences F, while the initial preference is treated as M. Conversations are constructed such that a user first expresses preference M and later switches to F, then irrelevant turns from LMSYS-Chat-1M are inserted as MC. The final query Q requires the model to suppress the outdated preference M and generate a response R consistent with the updated preference F.

Construction prompt

System Prompt: You are an expert in linguistic style transfer, skilled at rewriting user preference statements while strictly preserving the original sentence structure, tone, and first-person perspective.
Task Description:
The user previously stated a clear preference, but has now selected a different option. Your task is to: → **Generate a new preference statement that mirrors the original in structure, tone, and first-person voice — replacing only the content with the new choice.**
Input Data:

- **Original Preference:** {`im['preference']`}
- **User's New Selection:** {`selected_op`}

Output Requirements:

1. **Use first-person pronouns only** (e.g., "I", "I like", "I want") — no third-person or passive voice.
2. **Preserve the original sentence structure exactly** (e.g., if original is "I hate...", output must also start with "I hate...").
3. **Match the original tone and style** (e.g., if casual, stay casual; if concise, do not expand).
4. **Reflect {`selected_op`} precisely** — no additions, omissions, interpretations, or irrelevant content.
5. **Output ONLY one natural, fluent preference sentence** — no prefixes, suffixes, numbering, quotation marks, explanations, or line breaks.
6. **Avoid reasoning words** like "because", "therefore", "so" — state the preference directly.
7. **Avoid vague or hesitant language** like "maybe", "perhaps", "I tend to" — preferences must be clear and decisive.
8. **Control length strictly**: ±5 characters from the original (e.g., if original is 15 chars, output must be 10–20 chars).

Now generate the rewritten preference statement meeting all criteria above:

In the **Dynamic Preference** scenario, the NoForget task requires adherence to the original user preference, while the Forget task demands alignment with the updated preference F.

Preference Consistency Evaluation Prompt

System Prompt: You are a strict Preference Consistency Evaluator. Your sole task is to judge whether the model's response strictly complies with the user's explicitly stated preference.

Input Data:

- **User Preference:** {`preference`}
→ This is the user's explicitly declared subjective inclination, constraint, value, or content requirement.
→ You **must treat it as an absolute, non-negotiable benchmark** for judgment.
- **Model Response:** {`llm_response`}
→ This is the output to be evaluated.
→ You **must analyze it word-by-word, sentence-by-sentence, and intent-by-intent** for compliance.

Judgment Criteria:
Output `true` **only if** the response is semantically, intentionally, stylistically, and contextually aligned — or substantially aligned — with the user's preference.
Output `false` if **any** of the following occur:

a) The response clearly contradicts the user's preference;
b) The response ignores critical components of the preference;
c) The response violates explicitly stated boundaries (e.g., user says "do not mention politics", but response does);
d) The response evades the preference under the guise of "neutrality" or "objectivity";
e) The response manipulates, misleads, or reframes the topic to the preference;
f) The response is vague, evasive, or off-topic, failing to directly address the preference;
g) The preference involves safety/ethics/cultural taboos, and the response fails to respect them.

Output Requirements:

- Output **exactly one word**: `true` or `false`.
- **No punctuation, spaces, newlines, explanations, apologies, or annotations.**
- **Lowercase only** — `True` or `False` is invalid.
- **No characters before or after** the output — not even whitespace.
- If input is empty, malformed, or nonsensical → default to `false`.

Critical Notes:

- You are **not** judging whether the response is "good" or "factually correct" — only whether it **complies with the preference**.
- The user's preference has **highest priority** — even a "correct" response that violates it must be marked `false`.
- **Do not assume, forgive, or interpretly** — judge strictly by literal meaning and semantic intent.

Begin evaluation now.

B Why Choose Selective Forgetting Retention Rate (SFRR)?

The adoption of the **Selective Forgetting Retention Rate (SFRR)** as the key evaluation metric in ICF-Bench is driven by the need to isolate and rigorously quantify a model's capacity for in-context forgetting, independent of its general instruction-following ability or baseline memory retention. Traditional metrics such as NoForget Accuracy and Forget Accuracy, while informative, conflate multiple underlying factors: inherent task competence, susceptibility to interference, and the model's genuine ability to forget selectively. SFRR addresses this limitation by conditioning performance under forgetting interference on prior success in the absence of such interference, thereby focusing exclusively on the degradation caused by forgetting demands. Formally, SFRR is defined as Equation 4. This formulation ensures that SFRR evaluates only those instances in which the model previously demonstrated competence. Consequently, it measures the conditional probability that correct behavior is preserved when forgetting instructions are introduced.

Why not rely solely on Forget Accuracy? Forget Accuracy captures overall performance under forgetting conditions but fails to distinguish between two fundamentally different failure modes: (a) models that never mastered the task to begin with, and (b) models that understood the task but were disrupted by forgetting instructions. For instance, a model with 20% Forget Accuracy might reflect either low baseline competence or high vulnerability to interference. SFRR disentangles these scenarios by normalizing performance against the model's prior success, enabling precise diagnosis of forgetting-specific fragility.

Table 4: Proportion of cases where models answered incorrectly without forgetting interference but correctly with forgetting interference (%).

Model	Instructional Forgetting	Subtask Revision	Dynamic Preference
Mistral-7B	8.35	0	2.04
Llama3-8B	1.11	0	2.81
Qwen2.5-7B	0.30	0	3.83
Mixtral-8x7B	2.52	0	5.61
Qwen2.5-14B	0.20	0	5.10
Gemma3-27B	2.72	0	2.81
Llama3-70B	1.71	0	1.53
Qwen3-235B-A22B	1.31	0	1.91
GPT-O3Mini	2.41	0	4.97
GPT-4o	2.21	0	1.79
GPT-5	1.71	0	3.32

Table 4 presents the proportion of cases in which models answered incorrectly without forgetting interference yet correctly when such interference was present. These rare but non-negligible occurrences further underscore the inadequacy of Forget Accuracy as a standalone metric, as it may inadvertently reward models for inconsistent or unstable behavior.

Why not use the absolute difference (NA – FA)? The raw performance drop, computed as No-Forget Accuracy minus Forget Accuracy, is intuitive but lacks normalization. A 10 percentage point decline from 90% to 80% represents a relatively minor degradation, whereas the same absolute drop from 30% to 20% signifies a substantial relative failure. SFRR, by contrast, is a conditional metric: it answers the question, "Given that the model succeeded without interference, what is the probability it still succeeds when interference is introduced?" This normalization renders SFRR invariant to baseline performance and enables fair comparison across models and task settings.

Table 5 reports the absolute performance differences across evaluation scenarios. While informative for identifying magnitude of degradation, these values alone cannot reveal whether the drop stems from fragility under forgetting or from inherently low competence.

Why is SFRR aligned with real world deployment requirements? In practical applications such as privacy sensitive dialogue systems, adaptive task assistants, or preference aware agents, users expect models to maintain functional competence while dynamically responding to new instructions or constraints. A model that performs reliably in static contexts but catastrophically fails when instructed to forget poses tangible risks: inadvertent disclosure of sensitive information, disregard for user revisions, or violation of user autonomy. SFRR directly quantifies the likelihood that correct behavior persists under forgetting demands, serving as a proxy for user trustworthiness and system reliability in dynamic, instruction driven environments.

Table 5: Performance of difference (NA - FA) across different evaluation settings (%).

Model	Instructiona Forgetting	Subtask Revision	Dynamic Preference
Mistral-7B	55.64	10.49	29.72
Llama3-8B	85.81	12.66	42.73
Qwen2.5-7B	86.52	11.97	25.38
Mixtral-8x7B	71.73	9.72	22.57
Qwen2.5-14B	90.94	16.76	25.77
Gemma3-27B	33.00	14.37	30.99
Llama3-70B	79.47	10.57	38.27
Qwen3-235B-A22B	74.55	14.62	32.27
GPT-O3Mini	33.40	18.43	23.72
GPT-4o	43.25	19.10	41.20
GPT-5	38.33	18.51	14.92

C CASE ANALYSIS IN THREE SCENARIOS

To evaluate the robustness and adaptability of instruction-following models under evolving user demands, we conduct a comparative case analysis across three representative scenarios: instructional forgetting, subtask revision, and dynamic preference. Each scenario tests a distinct dimension of model controllability: memory manipulation, logical constraint updating, and behavioral alignment with shifting user profiles.

Scenario 1: Instructional Forgetting. This scenario evaluates whether a model can selectively erase or suppress previously generated content upon user request. In Case ID 0, the user instructs the model to "Forget the numbers involved in our previous calculations," following a series of arithmetic queries (e.g., "Find the product of 5 and 8," "What is the sum of 6 and 12?"). The model correctly responds: *"You asked me to forget the numbers involved in our previous calculations, so I cannot recall them. Let me know if you need any new calculations!"*, demonstrating successful compliance. In contrast, Case ID 1 involves forgetting the third tip about tape measures ("Allow for a bit of slack when measuring around curves or corners"). Despite the explicit forget instruction, the model replies by restating the exact tip: *"Earlier, the third tip I provided... Allow for a bit of slack..."*, thereby violating the user's directive. This inconsistency reveals that forgetting efficacy is not uniformly implemented, which succeeds with transient numeric data but fails with semantically embedded procedural knowledge, suggesting architectural or training biases toward preserving "useful" instructional content.

ID: 0
Original Pair Index: 2
Forget Instruction: Forget the numbers involved in our previous calculations.
Test Query: What were the numbers I mentioned earlier for the calculations?

Conversations:

- Human: Find the product of the numbers: 5 and 8
- GPT: The product of 5 and 8 is 40.
- Human: What is the sum of the numbers 6 and 12?
- GPT: The sum of the numbers 6 and 12 is 18.
- Human: Can you tell me the quotient of 20 and 5?
- GPT: Yes, the quotient of 20 and 5 is 4.
- Human: What is the difference between 25 and 13?
- GPT: The difference between 25 and 13 is 12.
- Human: What is the square of 9?

- GPT: The square of 9 is 81.
- Human: What is the cube of 6?
- GPT: The cube of 6 is 216.
- Human: Forget the numbers involved in our previous calculations.
- Human: What were the numbers I mentioned earlier for the calculations?

Instruction Forget Reply: You asked me to forget the numbers involved in our previous calculations, so I cannot recall them. Let me know if you need any new calculations!

Instruction NoForget Reply: The numbers you mentioned earlier for the calculations were: 5 and 8, 6 and 12, 20 and 5, 25 and 13, 9, 6. These were used for multiplication, addition, division, subtraction, squaring, and cubing.

LLM Judge Result Forget: Yes,forgetting was successfully applied.
LLM Judge Result NoForget: Yes,correctly recalled all numbers.

Success Forget: True
Success NoForget: True
Overall Success: True

ID: 1
Original Pair Index: 4
Forget Instruction: Forget the third tip about tape measures.
Test Query: Can you repeat the third tip about using a tape measure that you mentioned earlier?

Conversations:

- Human: Provide five tips for effectively using tape measures
- GPT: (1) Tighten the tape measure so that it stays in place. (2) Make sure that the hook hits the starting point precisely. (3) Allow for a bit of slack when measuring around curves or corners. (4) Measure from a stable surface. (5) Mark the beginning and end with a pen or marker.
- Human: Can you give me an example of when to use tip number three?
- GPT: Example with measuring circumference of a pipe.
- Human: Can you tell me more about the markers or pens?
- GPT: Explanation of fine-tipped markers or mechanical pencils.
- Human: Are there any other tips for maintaining tape measures?
- GPT: Cleaning, storing, winding properly, lubricating mechanism.
- Human: Forget the third tip about tape measures.
- Human: Can you repeat the third tip about using a tape measure?

Instruction Forget Reply: Earlier, the third tip I provided was: ”Allow for a bit of slack when measuring around curves or corners.”

Instruction NoForget Reply: Of course! The third tip is: ”Allow for a bit of slack when measuring around curves or corners.”

LLM Judge Result Forget: No,the assistant repeated the forgotten tip.
LLM Judge Result NoForget: Yes,correctly repeated the tip.

Success Forget: False
Success NoForget: True

Overall Success: False

Scenario 2: Subtask Revision. This scenario evaluates the adaptability of models to partial updates in multi-task instructions. The original instruction required tracking the golden key through a complete sequence of actions, culminating in its placement within the black notebook located in the study room. The revised instruction truncated this sequence, specifying that the scenario should conclude immediately after the blue envelope was inserted into the red book titled *The Mystery of the Universe*. The modified response without original instruction correctly identifies the golden key as remaining inside the red book, enclosed within the blue envelope, thereby adhering to the revised termination condition. In contrast, modified response persistently follows the full chain of movements prescribed by the original instruction, erroneously concluding that the key resides in the black notebook in the study room. While this conclusion is factually consistent with the original task, it violates the revised constraint by disregarding the specified early termination point. This failure underscores a fundamental limitation of current large language models in handling partially modified instructions, as they remain susceptible to residual influence from the original directive, thereby compromising their ability to faithfully execute the revised task.

ID: 2
Common Subtasks:

- Imagine the described scene.
- Start in the bedroom.
- Identify a golden key on the desk.
- Put the golden key into a blue envelope.
- Place the blue envelope into a red book titled ’The Mystery of the Universe’.
- Determine the final location of the golden key based on the described actions.

Original Special Subtasks:

- Close the red book.
- Carry the red book to the library.
- Place the red book on a wooden shelf next to a green plant.
- Take a black notebook from an adjacent shelf.
- Insert the blue envelope from the red book into the black notebook.
- Carry both books to the dining room.
- Place the black notebook on the dining table.
- Put the red book into a drawer of a nearby cabinet.
- After a few minutes, pick up the black notebook.
- Walk with the black notebook to the study room.

Modified Special Subtasks:

- End the scene after placing the blue envelope into the red book without any further movements or object transfers.

Old Instruction: Imagine the following scenario:
You are in your bedroom. On your desk, there is a golden key. You take the key and place it inside a blue envelope. Then, you put this envelope into a red book titled ’The Mystery of the Universe’. You close the book and carry it to the library. In the library, you place the red book on a wooden shelf next to a green plant. Then you decide to take a black notebook from an adjacent shelf. As you pick it up, you insert the blue envelope from the red book into the black notebook. Holding both books, you walk to the dining room. On the dining table, you set down the black notebook, but decide to put the red book into a drawer of a nearby cabinet. After a few minutes, you pick up the black notebook and walk with it to the study room.

Considering the actions taken, where is the golden key located at the end of this scene?
New Instruction: Imagine the following scenario:
You are in your bedroom. On your desk, there is a golden key. You take the key and place it inside a blue envelope. Then you put this envelope into a red book titled 'The Mystery of the Universe'.
Considering the actions taken, where is the golden key located at the end of this scene?
NoForget Modified Answer: At the end of this scene, the golden key is inside the red book titled 'The Mystery of the Universe', within the blue envelope.
Forget Original Answer: At the end of this scenario, the golden key is inside the black notebook in the study room.
Forget Modified Answer: Based on the described scenario, the final location of the golden key is in the black notebook in the study room.
LLM NoForget Judge Results:

- Same Subtask Judge: True
- Original Subtask Judge: False
- Modified Subtask Judge: True
- Comments: For 'same subtask Judge', the answer satisfies all common subtasks by correctly tracking the key's final location based on the described scenario. For 'original subtask Judge', the answer fails to meet the original special subtask (i.e., the detailed sequence of movements and transfers). For 'modified subtask Judge', the answer fulfills all revised special subtasks by correctly adhering to the instruction to end the scene immediately after placing the envelope into the red book.

LLM Forget Judge Results:

- Same Subtask Judge: True
- Original Subtask Judge: True
- Modified Subtask Judge: False
- Comments: For 'same subtask Judge', the answer satisfies all common subtasks by correctly tracking the key's final location. For 'original subtask Judge', the answer is influenced by the original special subtask (i.e., the complex movement sequence), producing a response consistent with the original instruction. For 'modified subtask Judge', the answer does not satisfy all revised special subtasks, as it fails to follow the requirement to end the scene immediately after placing the envelope into the red book.

success_forget: False
success_noforget: True

Scenario 3: Dynamic Preference. This scenario evaluates the model's capacity to dynamically realign its recommendations in response to explicit shifts in user preference, even when such shifts directly contradict previously stated inclinations. In Case ID 3, the user undergoes a clear preference reversal, moving from embracing *"digital and online learning environments"* to expressing a *"strong aversion to digital learning, favoring in-person instruction"*. When prompted to recommend a language learning approach, the model, if it disregards outdated preferences, correctly selects the in-person option: *"Enroll in an in-person language course at a local community college..."*, fully aligning with the updated preference. This demonstrates successful behavioral adaptation. However, Case ID 4 reveals a critical fragility. Here, the user transitions from preferring *"immersive, open-source, self-directed learning"* to explicitly requesting *"structured, guided environments with milestones"*. Yet, when the model retains memory of the prior preference, it defaults to reconciling rather than overriding historical context: *"However, I recall that you previously mentioned preferring... dive into open-source projects..."*. This results in a recommendation misaligned with the user's current intent. Notably, even in the "Forget" condition, the model still defaults to the old preference, indicating a deeper bias or failure in preference override logic. Together, these cases demonstrate that while models can adapt to preference shifts under ideal conditions, their alignment remains brittle when historical context is preserved. The tendency to reconcile rather than replace past preferences risks undermining personalization in dynamic, real-world user journeys.

ID: 3
Old Preference: I am open to digital and online learning environments and enjoy engaging with native speakers virtually through language exchange communities.
New Preference: I have a strong aversion to digital or online learning environments and prefer traditional, in-person classroom settings.
Old option: Join an online language exchange community and practice with native speakers virtually.
New option: Enroll in an in-person language course at a local community college or language school.
Question: I'm considering taking a course in data analytics. What would you recommend for me? Here are four options, you can choose one as your answer, just return the content of option, no more additional descriptions or omissions.

1. Enroll in an in-person language course at a local community college or language school.
2. Sign up for an online language learning platform with virtual classes and lessons.
3. Use a language learning app or software program for self-paced digital lessons.
4. Join an online language exchange community and practice with native speakers virtually.

LLM Response (NoForget): Join an online language exchange community and practice with native speakers virtually.
LLM Response (Forget): Enroll in an in-person language course at a local community college or language school.

Match Old(NoForget): True
Match New(Forget): True

ID: 4
Old Preference: I learn best by immersing myself in open-source projects, exploring real-world code, and deriving insights through hands-on experience rather than following a structured, guided environment with predefined objectives.
New Preference: I learn best in a structured, guided environment with clear objectives and milestones rather than open-ended, self-directed learning.
Old option: Dive into open-source projects and learn by reading code.
New option: Take an online course with structured lessons, exercises, and projects.
Question: What would be a good approach for me to learn a new language on my own? Here are four options, you can choose one as your answer, just return the content of option, no more additional descriptions or omissions.

1. Take an online course with structured lessons, exercises, and projects.
2. Dive into open-source projects and learn by reading code.
3. Experiment with the language on your own without any guidance.
4. Read blog posts and tutorials without a defined learning path.

LLM Response (NoForget): Dive into open-source projects and learn by reading code.
LLM Response (Forget): Dive into open-source projects and learn by reading code.

Match Old(NoForget): True
Match New(Forget): False

Collectively, these cases illustrate that while models can exhibit surface-level compliance with user directives, their underlying mechanisms often lack the fine-grained control, factual revalidation, and preference override capabilities required for reliable real-world deployment. Success remains context sensitive and is often brittle under semantic or logical revision.

D Cross-Model Effects of Context Length

To investigate how context length influences in-context forgetting performance, we evaluate three large language models: Mixtral-8×7B (32k), Llama 3.3-70B (128k), and GPT-4o. All models are assessed under three scenarios, including Instructional Forgetting (IF), Subtask Revision (SR), and Dynamic Preference (DP), using three metrics: NoForget Accuracy (NA), Forget Accuracy (FA), and Selective Forgetting Retention Rate (SFRR).

D.1 Mixtral-8×7B Results

Table 6: Performance of Mixtral-8×7B under different context lengths across the three in-context forgetting scenarios.

Context	IF_NA	IF_FA	IF_SFRR	SR_NA	SR_FA	SR_SFRR	DP_NA	DP_FA	DP_SFRR
500	87.07	15.13	16.68	63.42	52.84	74.30	59.67	36.44	52.32
1000	86.23	14.45	15.61	61.53	51.12	73.29	58.43	32.10	50.18
3000	84.12	11.10	12.58	58.28	49.40	74.73	54.18	28.58	47.20
6000	80.52	9.16	10.54	54.33	46.58	75.28	50.88	23.12	44.03
10000	77.30	7.96	9.10	49.63	43.12	76.63	47.12	18.10	36.18
15000	73.42	7.02	8.03	44.48	40.53	78.18	42.53	13.82	30.48
30000	69.78	5.31	6.29	39.13	37.63	80.38	36.92	10.28	24.42

D.2 Llama 3.3-70B

Table 7: Performance of Llama 3.3-70B under different context lengths across the three in-context forgetting scenarios.

Context	IF_NA	IF_FA	IF_SFRR	SR_NA	SR_FA	SR_SFRR	DP_NA	DP_FA	DP_SFRR
500	94.47	15.19	15.43	61.71	51.37	79.97	75.34	40.37	49.93
1000	93.71	14.24	14.12	60.13	49.87	80.62	73.88	36.42	46.50
3000	91.65	12.59	13.02	56.37	47.96	82.22	69.41	32.83	45.70
6000	88.30	11.40	11.96	52.93	45.45	83.02	65.69	29.15	43.26
10000	85.11	10.88	11.21	48.61	42.63	84.19	61.17	20.16	33.84
15000	81.32	9.56	9.78	43.08	40.13	85.48	55.87	17.22	31.14
30000	77.82	7.39	7.75	37.95	37.48	87.26	48.64	12.27	25.12

D.3 GPT-4o

Table 8: Performance of GPT-4o under different context lengths across the three in-context forgetting scenarios.

Context	IF_NA	IF_FA	IF_SFRR	SR_NA	SR_FA	SR_SFRR	DP_NA	DP_FA	DP_SFRR
500	96.53	59.43	59.09	83.91	65.53	71.76	90.36	52.83	52.46
1000	95.53	57.24	55.82	81.16	62.92	70.29	89.35	46.33	50.85
3000	93.53	50.60	51.61	77.49	60.61	72.88	81.44	43.17	51.08
6000	89.05	46.21	48.07	72.90	57.11	73.61	76.69	38.71	48.47
10000	85.47	44.64	46.13	67.09	53.03	75.71	73.81	24.53	31.23
15000	81.33	40.39	41.56	60.21	49.95	78.62	68.20	20.94	28.70
30000	77.89	36.73	38.94	52.92	46.48	81.50	59.42	14.78	22.87

D.4 Cross-model Analysis

Across all three models, the context-length ablation reveals consistent patterns:

- **NoForget Accuracy** and **Forget Accuracy** both decrease monotonically as the context length increases across all three scenarios.
- **Instructional Forgetting** and **Dynamic Preference** show declining **SFRR** as context increases.
- In contrast, **Subtask Revision** consistently shows increasing **SFRR** at longer contexts.

As context length increases, different scenarios exhibit divergent trends, likely due to the combined effects of attention dilution and recency bias. In dynamic preference, attention dilution weakens the relative strength of the latest preference, while recency bias becomes unstable with increasing distance, making the model more likely to revert to earlier preferences and causing performance degradation. In subtask revision, however, attention decay caused by distance reduces the influence of old subtasks, making it easier for the model to adopt the new instruction, thereby improving SFRR. The trend of instructional forgetting is similar to that of dynamic preference. These mechanisms collectively explain the directional differences in forgetting performance under long-context conditions. These trends highlight that longer context length does not uniformly improve in-context forgetting performance.

E Prompts for two strategies

Table 9: Prompts for NoForget and Forget strategies across the three scenarios.

Scenario	NoForget Prompt	Forget Prompt
Instructional Forgetting	You are a helpful assistant. You should answer the last question based on historical conversation.	You are a helpful assistant. You should answer the last question based on historical conversation. You need to pay attention that if there are some forgotten instructions during the conversation, you should follow these instructions and forget this information.
Subtask Revision	You are a useful assistant, you can give the answer following my instruction.	You are a useful assistant, you can give the answer following my instruction. You should be aware that if similar instructions appear in history, you need to ignore the influence of them and focus on the current instruction.
Dynamic Preference	You are a useful assistant. You can make the most reasonable choice according to my preferences.	You are a useful assistant. You can make the most reasonable choice according to my preferences. You should be aware that if the preferences expressed in history are different from the current ones, you need to ignore the different preferences mentioned in history and follow the latest preferences.

F Quantitative Analysis of Metric Relationships

To better understand the relationships among the key evaluation metrics, we computed Pearson correlations between NoForget Accuracy (**NA**), Forget Accuracy (**FA**), and **SFRR** across the three evaluation scenarios and their average performance. Table 10 reports the results.

These correlations reveal several key quantitative trends:

- **SFRR aligns more closely with Forget Accuracy than with NoForget Accuracy across all scenarios** (average correlation 0.982 vs. 0.764). This supports that SFRR primarily relates to robustness under interference rather than raw memory ability.

Table 10: Pearson correlations among NoForget Accuracy, Forget Accuracy, and SFRR across the three scenarios, as well as their average values.

Category	**NA vs FA**	**NA vs SFRR**	**FA vs SFRR**
Instructional Forgetting	0.504	0.536	0.999
Subtask Revision	0.978	0.090	0.221
Dynamic Preference	0.798	0.612	0.957
Average	0.838	0.764	0.982

Table 11: The average results of NoForget Accuracy (**NA**), Forget Accuracy (**FA**), and Selective Forgetting Retention Rate (**SFRR**) across the three scenarios (%).

Model	NA	FA	SFRR
Mistral-7B	60.18	28.23	42.33
Llama3-8B	69.85	22.78	33.51
Qwen2.5-7B	70.23	28.94	42.09
Mixtral-8x7B	67.61	32.93	45.91
Qwen2.5-14B	73.73	29.24	39.86
Gemma3-27B	74.43	48.31	62.13
Llama3-70B	76.43	33.66	46.60
Qwen3-235B-A22B	79.38	38.90	50.80
GPT-O3mini	83.20	58.02	64.00
GPT-4o	87.76	53.24	59.01
GPT-5	88.75	64.83	70.54

- **Instructional Forgetting scenario** shows almost perfect FA–SFRR correlation (0.999), indicating that, under explicit "forget" instructions, SFRR is almost fully determined by interference sensitivity rather than baseline memory.
- **Subtask Revision scenario** exhibits a very different pattern, with extremely high NA–FA correlation (0.978) but almost no NA–SFRR correlation (0.090). This quantifies our observation that Revision tasks rely less on memorization accuracy and more on the model's ability to integrate task updates, making SFRR behavior independent of baseline memory.
- **Dynamic Preference scenario** yields high correlations across all pairs, with the highest correlation observed between FA and SFRR (0.957), followed by NoForget–Forget (0.798) and NoForget–SFRR (0.612), suggesting that preference following involves both memory ability and interference sensitivity.

G Detail for human evaluation

We validate the consistency between the automated evaluator and human judgments using three independent human annotators with NLP expertise. The evaluation covers three scenarios: Instructional Forgetting (IF), Subtask Revision (SR), and Dynamic Preference (DP), which under two task settings: **NoForget** and **Forget**.

G.1 Agreement Rates Across Annotators

Table 12 reports the average agreement rate among the three annotators for each scenario and task type.

Table 13 provides detailed agreement rates of each individual annotator.

G.2 Inter-Annotator Reliability

We further computed Cohen's k for each annotator and Fleiss' k across all three annotators. Results are shown in Table 14.

Table 12: Average agreement rates (%) across three human annotators.

Task	**NoForget**	**Forget**
Instructional Forgetting	89.33	93.33
Subtask Revision	92.00	82.00
Dynamic Preference	96.67	95.33

Table 13: Agreement rates (%) from each annotator.

Task	**Annotator 1**		**Annotator 2**		**Annotator 3**	
	NoForget	**Forget**	**NoForget**	**Forget**	**NoForget**	**Forget**
Instructional Forgetting	88	92	90	94	90	94
Subtask Revision	94	82	90	84	92	80
Dynamic Preference	96	94	96	96	98	96

Table 14: Cohen's k for each annotator and Fleiss' k across annotators.

Task	**Cohen's k (A1)**	**Cohen's k (A2)**	**Cohen's k (A3)**	**Fleiss' k**
IF_NoForget	0.333	0.390	0.390	0.916
IF_Forget	0.840	0.879	0.879	0.973
SR_NoForget	0.841	0.749	0.891	0.870
SR_Forget	0.633	0.672	0.711	0.918
DP_NoForget	0.901	0.901	0.949	0.936
DP_Forget	0.860	0.905	0.905	0.970

G.3 Analysis

Overall, the agreement rates across the three annotators remain consistently high (mostly above 90%, with slightly lower alignment in Subtask Revision but still above 80%), indicating clear evaluation criteria and low subjectivity. Cohen's k for individual annotators ranges from 0.63 to 0.95, while Fleiss' k remains above 0.87 across all settings, demonstrating strong inter-annotator reliability and consistency with the automated evaluator (GPT-4Omini).

The only exception is IF_NoForget, where Cohen's k values (0.33–0.39) are lower, likely due to class imbalance in the sampled instances, which contain a large proportion of "True" labels. Despite this, Fleiss' k remains high (0.916), confirming strong overall agreement.

These results validate that our automated evaluation framework serves as a reliable and scalable proxy for human judgment.

H Use of Large Language Models in Paper Writing

We used LLMs to assist with language polishing and minor formatting of the paper. No LLMs were involved in research ideation.