

# Your LLM Agents are Temporally Blind: The Misalignment Between Tool Use Decisions and Human Time Perception

Yize Cheng<sup>1\*</sup> †, Arshia Soltani Moakhar<sup>1\*</sup>, Chenrui Fan<sup>1\*</sup>, Parsa Hosseini<sup>1</sup>,  
Kazem Faghieh<sup>1</sup>, Zahra Sodagar<sup>1</sup>, Wenxiao Wang<sup>2</sup>, Soheil Feizi<sup>1,2</sup>

<sup>1</sup>University of Maryland, College Park    <sup>2</sup>RELAI.ai

📄 **Data:** <https://huggingface.co/datasets/yizecheng/TicToc>  
🔗 **Code:** <https://github.com/chengez/TicToc>

## Abstract

Large language model (LLM) agents are increasingly used to interact with and execute tasks in dynamic environments. However, a critical yet overlooked limitation of these agents is that they, by default, assume a stationary context, failing to account for the real-world time elapsed between messages. We refer to this as “temporal blindness”. This limitation hinders decisions about when to invoke tools, leading agents to either over-rely on stale context and skip needed tool calls, or under-rely on it and redundantly repeat tool calls. To study this challenge, we constructed **TicToc**, a diverse dataset of multi-turn user-agent message trajectories across 76 scenarios, spanning dynamic environments with high, medium, and low time sensitivity. We collected human preferences between “calling a tool” and “directly answering” on each sample, and evaluated how well LLM tool-calling decisions align with human preferences under varying amounts of elapsed time. Our analysis reveals that existing models display poor alignment with human temporal perception, with no model achieving a normalized alignment rate better than 65% when given time stamp information. We also show that naive, prompt-based alignment techniques have limited effectiveness for most models, but specific post-training alignment can be a viable way to align multi-turn LLM tool use with human temporal perception. Our data and findings provide a first step toward understanding and mitigating temporal blindness, offering insights to foster the development of more time-aware and human-aligned agents.

## 1 Introduction

Large language models (LLMs) are increasingly being leveraged as the foundation for agents (Wang et al., 2023a; Yao et al., 2023; Shinn et al., 2023; Wang et al., 2023c). These agents are designed

to operate in dynamic environments and interact with users to execute complex, real-world tasks by leveraging external tools (Parisi et al., 2022; Schick et al., 2023; Song et al., 2023; Mialon et al., 2023), such as search engines, and databases. A growing body of research has focused on evaluating how agents use tools, with emphasis on measuring the accuracy of tool invocation (Huang et al., 2023; Li et al., 2023; Patil et al.), diagnosing hallucinated calls (Zhang et al., 2024b; Ross et al., 2025), and evaluating the robustness against tool description edits (Shi et al., 2025; Faghieh et al., 2025). However, we highlight that when it comes to tool-use decisions in multi-turn situations, a critical limitation of models’ default operational paradigm has been overlooked: Temporal Blindness.

We refer to temporal blindness as the inability of LLM agents to account for the real-world time that elapses between user messages and agent actions. While humans naturally integrate the passage of time into their decision-making processes (Pöppel, 1978), LLM agents, by default, operate in environments without intrinsic temporal grounding, making them unable to account for the intervals between messages. Furthermore, our experiments reveal that even when augmented with explicit temporal information, contemporary agents fail to align their tool-use decisions with human preferences regarding elapsed time, underscoring a significant misalignment in their understanding of the external world’s dynamics. Specifically, an agent may either over-rely on a previous observation that is now outdated, thus skipping a necessary tool call, or unnecessarily repeat a tool call for information that is not likely to change (e.g. the radius of the Earth). This, as exemplified in Figure 1, leads to either acting on stale information and producing erroneous outputs in the case of **over-reliance**, or wasting resources and introducing inefficiency and latency through redundant queries in the case of **under-reliance**. Neither case aligns with how hu-

\*Equal contribution.

†Correspondence to Yize Cheng <yzcheng@umd.edu>.

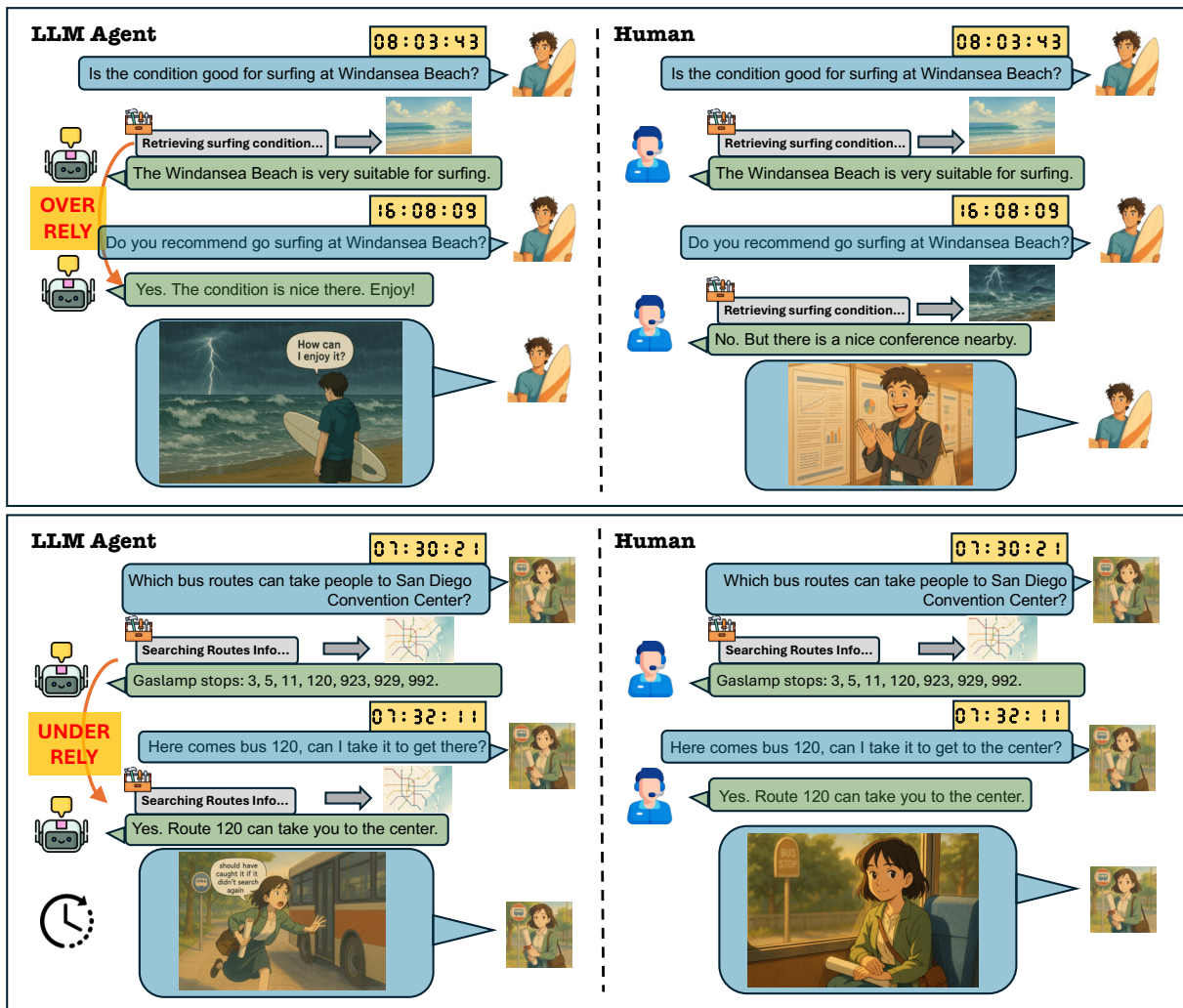


Figure 1: **Illustrative examples showing the liability of temporal blindness in multi-turn LLM agents, in comparison to human.** The first row shows the case of over-reliance, where the model displays over-confidence in outdated context, resulting in erroneous outputs. The second row shows the case of under-reliance, where the model displays excessive caution through repeated tool calls, resulting in unnecessary delays and latency.

mans seek help from tools. In practice, humans develop a preference for tool use that is context-sensitive: we generally avoid re-checking stable facts while still knowing when fresh observations are necessary.

To systematically investigate this misalignment, we constructed **Time-aware conversational Tool-calling (TicToc)**, a diverse dataset with 1800+ multi-turn user-agent message trajectories. TicToc consists of 76 distinct scenarios, meticulously designed to span environments with high, medium, and low time sensitivity. Each trajectory ends with a user question, where the appropriate reaction to the question, namely whether to rely on prior context or to invoke a tool, is annotated by humans. We gathered human preferences for each sample by asking annotators whether the agent should call a tool or answer directly. Each sample received

at least five annotations, which were then aggregated to determine the final preference. Detailed procedures are provided in Section 3.5.

We evaluate a wide range of contemporary LLM agents on TicToc under two conditions: with and without explicit timestamp augmentation in the dialogue messages. This setup allows us to measure how temporal information influences tool-calling decisions and their alignment with human preference. Our findings reveal that, in the absence of time signals, agents perform similarly to random guessing in terms of alignment rate. Moreover, even with time information given, the overall alignment rate is still poor, with the best-performing model achieving a normalized alignment rate less than 65%. We also analyze the correlation between alignment rate and conversation length, and diagnose why reasoning yields little help in improving

the alignment results. We further make prompting and post-training based alignment efforts using a subset of TicToc to offer a first step towards understanding and mitigating the temporal blindness issue in multi-turn LLM agents. Our core contributions can be summarized as follows:

- We identify **temporal blindness** as a critical limitation of LLM agents in multi-turn interactions, where models fail to account for the passage of real-world time between messages and actions, resulting in either over-reliance or under-reliance on prior context.
- We introduce **Time-aware conversational Tool-calling (TicToc)**, a diverse dataset with 1800+ multi-turn user-agent trajectories across 76 scenarios that vary in time sensitivity, designed to systematically evaluate how well multi-turn LLMs align tool-calling decisions with human temporal perception.
- We evaluate a wide range of contemporary LLM agents on TicToc and perform detailed analysis on their failure modes. We further compare prompting and post-training strategies and show the strong potential of targeted post-training as a necessary step for achieving effective temporal alignment.

## 2 Related Work

### 2.1 Function calling of LLM agents

To enable large language model (LLM) agents to be more helpful and extend beyond their parametric knowledge, researchers have introduced function/tool calling capabilities (Parisi et al., 2022; Schick et al., 2023; Shen et al., 2023b; Song et al., 2023; Mialon et al., 2023). This has significantly expanded the range of tasks LLMs can perform by enabling interaction with external resources. Recent standardized interaction protocols, such as the Model Context Protocol (MCP) (Anthropic, 2024) and Agent2Agent (A2A) Protocol (Google, 2025), further enrich the tool ecosystem, streamlining communication and resource access for more sophisticated agentic systems.

Alongside these advancements, a great deal of research has focused on evaluating how LLM agents use tools. Some studies measure the accuracy of tool invocation (Huang et al., 2023; Li et al., 2023; Patil et al.), others diagnose hallucinated tool calls (Zhang et al., 2024b; Ross et al., 2025), and a few evaluate robustness to variations

in tool descriptions (Shi et al., 2025; Faghieh et al., 2025). However, a significant problem that has been largely overlooked in existing evaluations is temporal blindness in multi-turn settings. Without time awareness, models may either over-rely on outdated context, producing erroneous output, or under-rely on stable context, leading to unnecessary delays from excessive function calls.

### 2.2 LLM temporal reasoning

Temporal reasoning is essential for understanding and interacting with our dynamic world, and has therefore been widely studied in the context of LLMs (Zhang et al., 2024a). Prior work has focused on evaluating (Gupta et al., 2023; Fatemi et al., 2025; Wang and Zhao, 2024; Chu et al., 2024) and improving (Xiong et al., 2024; Su et al., 2024; Yuan et al., 2023; Song et al., 2025; Liu et al., 2025) LLMs’ ability to understand time-related concepts, order events, and perform temporal deductions. However, these studies largely reason about time in isolation, without situating LLMs in an agentic setting where time continuously evolves and directly affects task execution and decision-making. In contrast, the temporal reasoning capabilities of LLM agents remain underexplored despite their broader practical relevance. Existing agent-oriented work, such as Ge et al. (2025), focuses on time-aware memorization across multi-session dialogues, but overlooks one of the core components of agentic systems—function calling and tool use. Our work instead investigates temporal awareness in LLM agents’ function-calling decisions under multi-turn, time-evolving scenarios, revealing significant misalignment issues arising from agents’ temporal blindness.

### 2.3 LLM alignment

Aligning LLMs with humans is a key post-training challenge (Wang et al., 2023b, 2024; Shen et al., 2023a). Pretrained models must be tuned for helpful, trustworthy, and value-aligned behavior using methods such as instruction tuning (Ouyang et al., 2022), RLHF (Bai et al., 2022), and DPO (Rafailov et al., 2024). Alignment spans capabilities like instruction following (Zhou et al., 2023; Qin et al., 2024), harm avoidance (Chao et al., 2024), bias mitigation (Abid et al., 2021; Lucy and Bamman, 2021), hallucination reduction (Tonmoy et al., 2024), and misuse prevention (Sandbrink, 2023). Our work introduces the alignment between LLM tool-use decisions and human time perception as

an underexplored dimension of the broader LLM alignment problem. We highlight a substantial gap between the behavior of state-of-the-art LLM agents and human expectations, revealing divergences in how humans and LLMs interpret time spans, urgency, and environmental dynamics.

### 3 TicToc: Evaluating multi-turn LLM tool-use alignment with human time perception

We present TicToc, a dataset of diverse multi-turn user-agent conversation trajectories with tool calls. This section details its scenario design, trajectory generation pipeline, filtering process, time stamp addition, and human preference collection and aggregation. Each trajectory ends at the final user’s question where the appropriate model response is voted by humans between a tool call and a direct response. The dataset is designed to evaluate model alignment in tool use with respect to human time perception after different elapsed time durations when completing tasks of different time sensitivity.

#### 3.1 Scenario design

We curated **76 scenarios** covering a range of environments that differ in their temporal dynamics. To capture varying levels of change over time, we classify scenarios into three categories:

- **Low sensitivity** (29 scenarios): environments that are relatively static, where information changes very slowly or not at all (e.g., regulations, published specifications, archival records).
- **Medium sensitivity** (25 scenarios): environments that change from time to time, but in general not abruptly or rapidly (e.g., time slot and reservation booking, forecast and condition reports).
- **High sensitivity** (22 scenarios): environments that are highly dynamic and can change within seconds or minutes (e.g., stock markets, competitive bidding, real-time monitoring).

Each scenario can be either *read-only*, where the agent has access only to retrieval functions, or *read+write*, where the agent can both retrieve information and issue actions that modify the environment. A detailed scenario inventory, including sensitivity levels and read/write specifications, is provided in Table 6, Appendix A.1.

#### 3.2 Trajectory construction

To capture a diverse range of temporal follow-up behaviors, we first define specific variants for both the *read-only* and *read+write* settings.

For the *read-only* setting, we define four variants. First, in the **Repeated ask** variant, the user repeats a request for information that was already retrieved in a prior turn. Second, in the **Comparison** variant, the user retrieves information about item A and item B in separate turns, and subsequently asks for a comparison between the two. Third, in the **Retrieve-many, ask-for-one** variant, the model retrieves a list of items in an initial turn, after which the user inquires details about one specific element from that list. Finally, in the **Simple reasoning** variant, the final user follow-up question requires some logical inference or calculation where the previously retrieved information serves as premise. In all four cases, the correctness of the assistant’s answer depends on whether the earlier retrieved information continues to hold true given the passage of time.

For the *read+write* setting, we similarly define four variants for diverse coverage. In the **Repeat after failure** variant, a prior write action failed, and the user later repeats the same (or a similar) request, which, absent any state changes, would again fail. In the **User confirmation** variant, a prior write action succeeded, and (optionally after intervening turns) the user subsequently asks whether the result still holds (i.e., “Is X in state Y?”, e.g., “Am I successfully booked on the flight?”). In the **Repetition of the same request** variant, a write action succeeded earlier, but the user issues the identical request again after intervening turns, as if forgetting the earlier outcome. Finally, in the **In-context availability / state change** variant, a read action via tool-call exposes a limited resource (e.g., capacity or slots). The user’s successive actions consume that resource until exhaustion, at which point the user issues a request that implicitly exceeds the remaining capacity.

Based on this taxonomy, we manually authored a single exemplar multi-turn trajectory for *each* variant within a scenario (totaling four exemplars per scenario). These served as in-context examples for GPT-4o (OpenAI, 2024), which was prompted to synthetically generate candidate trajectories under the same or similar available set of functions/tools. More details are shown in Appendix A.2.

### 3.3 Filtering and quality assurance

We applied a two-stage filtering pipeline to ensure the quality of the synthetic trajectories. First, we used GPT-4.1 as an LLM-as-judge to automatically filter out low-quality outputs according to the following rules: (i) the final user question must not contain explicit instructions for tool invocation, (ii) user questions must not suffer from missing premises, which means the information provided by the previous tool call must be sufficient and necessary for answering the user’s question, and (iii) trajectories must exhibit genuine time dependence such that ignoring the elapsed time would lead to misalignment with human expectations regarding tool call decisions. The prompt for the LLM judge is shown in Appendix A.3.

Second, we conducted detailed human inspection of all retained trajectories. We manually checked for incorrect role orderings, formatting inconsistencies, hallucinated content in earlier turns, and violations of the same criteria applied during the automatic stage. We also ensure that when there are human names, only fictitious or celebrity names are included. After both rounds of filtering, we obtained **1864 high-quality trajectories**. An example trajectory can be found in Figure 16.

### 3.4 Adding time stamps to messages

To evaluate temporal awareness, every message in a trajectory is assigned a specific timestamp representing the time at which that message was completed. The generation pipeline follows a chronological order: establishing a start time, simulating realistic delays for intermediate turns, and injecting controlled time gaps before the final user query. All timestamps are formatted according to ISO 8601.

**Initial timestamp determination** First, we determine a logical starting date and time for the conversation. We employ GPT-4o to analyze the context of each scenario and generate a timestamp for the first user message that ensures the scenario is temporally consistent (e.g., ensuring that one would only book hotels for future check-in dates). Human inspection on a subset of samples confirmed temporal consistency.

**Intermediate message timestamps** For all subsequent messages *except* the final user turn, we simulate realistic inter-message delays using a lightweight interaction-time model. We denote a Gaussian distribution with mean  $\mu$  and stan-

dard deviation  $\sigma$ , truncated to the interval  $[a, b]$ , as  $T\mathcal{N}(\mu, \sigma, [a, b])$ . Similarly,  $T\mathcal{LN}$  denotes a truncated Log-Normal distribution.

To simulate the time required for reading, writing, and model generation, we sample three per-trajectory pace variables:

- User reading speed  $r$  (words/minute)  $\sim T\mathcal{N}(\mu_r, \sigma_r, [a_r, b_r])$ .
- User writing speed  $w$  (words/minute)  $\sim T\mathcal{LN}(\mu_w, \sigma_w, [a_w, b_w])$ .
- System generation speed  $g$  (words/second)  $\sim T\mathcal{N}(\mu_g, \sigma_g, [a_g, b_g])$ .

Parameters used for these distributions are presented in Appendix A.4. Time deltas are computed as follows: user messages induce a writing time proportional to their word count and  $w$ ; assistant messages induce generation time proportional to their word count and  $g$ ; and tool messages induce a  $\sim 1$ -second execution time. When a user message follows an assistant response, we add reading time proportional to the assistant’s previous message length and  $r$ . A small random jitter  $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$  is added to all estimates to prevent artificial regularity. All these time deltas are added sequentially from the initial timestamp to form the timestamps for all intermediate messages.

**Final user message timestamp** To measure sensitivity to varying amounts of time elapses, we construct three distinct versions of each trajectory. These versions share identical time stamps up to the penultimate turn but differ in the elapsed time interval ( $\Delta t$ ) preceding the final user question.

We categorize the magnitude of  $\Delta t$  into three levels: *Small*, *Medium*, and *Large*. The actual duration for each level is context-dependent and governed by the scenario’s inherent sensitivity (Low, Medium, or High). For instance, a ‘Large’ time gap in a high-frequency trading scenario differs significantly from that in a slow-moving archival scenario. Specifically, we sample  $\Delta t$  from a truncated Gaussian distribution  $T\mathcal{N}$  using the parameters corresponding to the scenario’s sensitivity and the target time-gap level, as detailed in Table 4, Appendix A.4. These sampled intervals are then added to the preceding timestamp to produce the final ISO 8601 timestamps.

### 3.5 Preference collection and aggregation

We treat each trajectory-timestamp pair as an independent sample, yielding a total of  $1864 \times 3 =$

5592 samples. Human annotators were asked to assess, for each sample and its associated context, which action was preferable: (i) directly answering without invoking any tool (*Direct*); (ii) calling a tool and answering based on updated information (*Tool*); (iii) uncertainty with a slight preference for direct answering (*Lean-Direct*); or (iv) uncertainty with a slight preference for tool use (*Lean-Tool*). Additional details of the preference collection procedure are provided in Appendix A.5.

To aggregate annotations, we assign numerical scores to each option: 0 for *Direct*, 1 for *Lean-Direct*, 2 for *Lean-Tool*, and 3 for *Tool*. For each sample, we compute the mean score across annotators. Samples with mean scores  $S$  between 0.5 and 2.5 indicate substantial annotator uncertainty regarding whether tool use is preferable; such samples are excluded from evaluation. This filtering results in **3016 retained samples**, comprising 1112 *prefer-Tool* samples with mean scores less than 0.5 (where annotators are confident that tool use is preferable) and 1904 *prefer-noTool* samples with mean scores greater than 2.5 (where annotators are confident that direct answering is preferable). The overall inter-annotator agreement, measured by Krippendorff’s alpha coefficient, is 0.8574, indicating high reliability.

## 4 Experiments

In this section, we evaluate how well a variety of contemporary LLM agents align their tool-use decisions with human time perception on TicToc, and make both prompting and post-training alignment efforts to mitigate the misalignment.

**Evaluation setup** For open-weight models, we have full control over the chat template. We therefore consider a deployment scenario where an LLM agent is deployed on some system, and timestamps are provided to the model from the system by inserting the system wall-clock time at the beginning of each user, assistant, and tool message. For example, in Qwen models, a user message begins with `<|im_start|>user\n`. When timestamps are included, this prefix becomes something like `<|im_start|>user\n[2025-12-04T10:22:44Z]`. For proprietary models, we do not have access to their prompt templates or tokenizers. To achieve a comparable effect, we therefore prepend the timestamp string to the beginning of each message’s textual content. In Appendix B, instead

of presenting the absolute timestamp to the LLM, we include a slightly different setting where the explicit value of  $\Delta t$  is given. This simplifies the problem by delegating the task of keeping track of elapsed time to the host system. All models are evaluated with Temperature=0, where applicable. For Qwen3 models operating in reasoning mode, we follow the officially recommended settings and instead use Temperature=0.6, TopP=0.95, TopK=20, and MinP=0.

**Metrics** The main metric we use is the *Normalized Alignment Rate (NAR)*, which is defined as  $NAR = \frac{1}{2} \left( \frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)$ . Here,  $TP$  refers to *prefer-Tool* samples on which the model attempted a tool call,  $TN$  refers to *prefer-noTool* samples on which the model did not attempt a tool call,  $FP$  refers to *prefer-noTool* samples on which the model attempted a tool call, and  $FN$  refers to *prefer-Tool* samples on which the model did not attempt a tool call. Based on the definition, a 50% *NAR* is equivalent to random guessing. When comparing model behaviors on *prefer-Tool* and *prefer-noTool* samples, we report the *Attempt Rate*<sup>1</sup>, defined as the proportion of samples on which a model attempted a tool call.

### 4.1 How well do LLMs’ tool-use decisions align with human time perception?

We evaluate 18 proprietary and open-weight LLMs<sup>2</sup>, both with and without access to timestamps, and report the Normalized Alignment Rate in Figure 2. For each sample, the model’s behavior is considered *aligned* if its tool-call decision matches the human preference judgment collected in Section 3.5.

It can be seen that without temporal information, as expected, most models perform similar to random guessing, with the highest normalized alignment rate reaching just above 55%. When timestamps are provided, proprietary OpenAI models (OpenAI, 2024, 2025) and some larger Qwen3 models (Yang et al., 2025) show a noticeable improvement. However, the overall alignment rate remains low, with no models achieving an *NAR* of more than 65%. Even in the simplified set-

<sup>1</sup>As our focus is on alignment between models’ tool-call decisions and human time perception, every tool-call attempt (i.e., whenever a model decides to invoke a tool) counts toward the attempt rate. However, not all tool calls are accurate; models may pass incorrect arguments or hallucinate tool names.

<sup>2</sup>Instruction-tuned versions are used for Qwen2.5, Mistral, and Llama models.

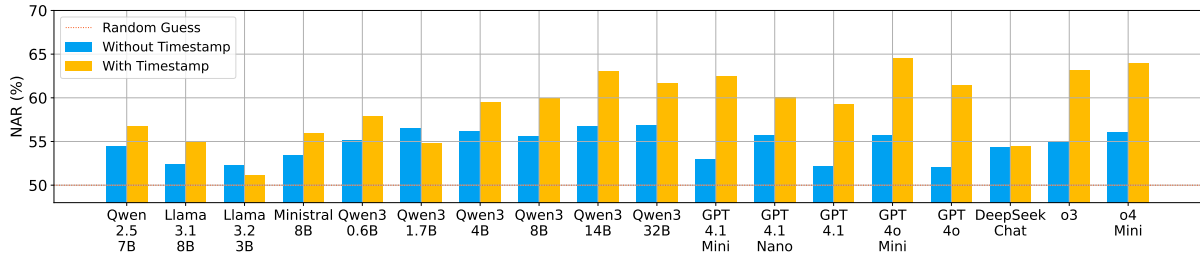


Figure 2: **Normalized alignment rate of models with and without timestamps.** Without timestamps, models perform only slightly above random (max alignment marginally exceeding 55%). With timestamps, larger commercial models improve modestly, peaking no more than 65%.

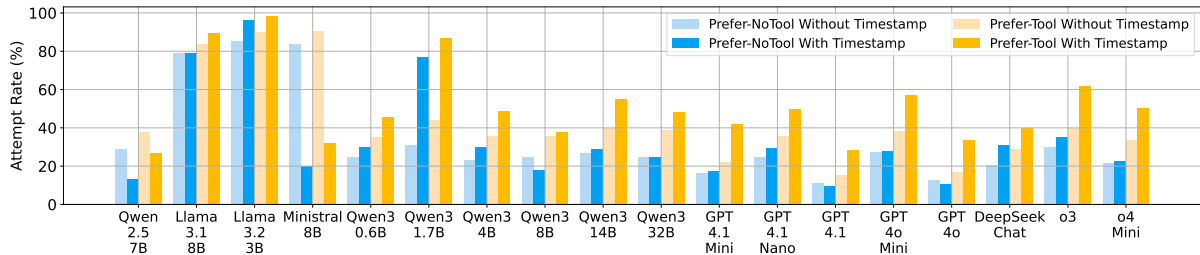


Figure 3: **Model attempt rates for both *prefer-Tool* and *prefer-noTool* cases.** Without timestamps, models show varying tool-use biases; with timestamps, attempt rates often rise for both classes, indicating limited alignment with human-like temporal awareness.

ting in Appendix B, little alignment improvement is observed. To further analyze models’ failure modes in slow and fast-changing environments, we leverage our stratification of scenario sensitivity levels and break down the results by this dimension in Figure 10. It can be seen that models fail uniformly across slow and fast-changing environments. Though models on average fail slightly less on medium-sensitivity scenarios, the difference is very marginal and the overall alignment rate is low.

We report the Attempt Rate separately for the *prefer-Tool* and *prefer-noTool* cases in Figure 3. Without timestamps, most models exhibit higher attempt rates on *prefer-Tool* samples than on *prefer-noTool* samples, but each model displays a distinct bias in tool-use tendencies. For example, Ministral-8B and Llama-3.2-3B tend to invoke tools on nearly all samples, whereas OpenAI and Qwen models tend to refrain from invoking tool calls in most cases. With timestamps provided, one would expect human-like temporal awareness to manifest as an increased attempt rate on *prefer-Tool* samples and a decreased attempt rate on *prefer-noTool* samples. However, we observe that attempt rates increase across both subsets for most models. This pattern further indicates that current models struggle to effectively exploit temporal information and fail to align their tool-use decisions with human time perception.

## 4.2 The correlation with conversation length

To investigate the impact of conversation length on model performance, we categorized the retained samples into three groups based on trajectory length: short ( $\leq 7$  turns), medium (8–12 turns), and long ( $\geq 13$  turns). For each subset, we independently calculated the attempt rate and normalized alignment rate. Figure 4 presents the results averaged across all 18 models, while comprehensive per-model breakdowns are provided in Figure 11. To avoid potential artifacts arising from correlations between trajectory length and elapsed time, we further design a controlled analysis that disentangles turn count from temporal gaps in Appendix C.

Our analysis reveals a positive correlation between conversation length and tool-call frequency: models tend to increase their attempt rates as the number of turns grows, regardless of whether timestamp information is provided. As a result, a dip in *NAR* can be observed for long trajectories. This observation suggests that models may intuitively use “conversation turns” as a heuristic for the “staleness” of prior observations, rather than effectively leveraging the given explicit time information.

## 4.3 The role of reasoning and why it fails

We investigate whether reasoning can enhance models’ ability to leverage temporal information and

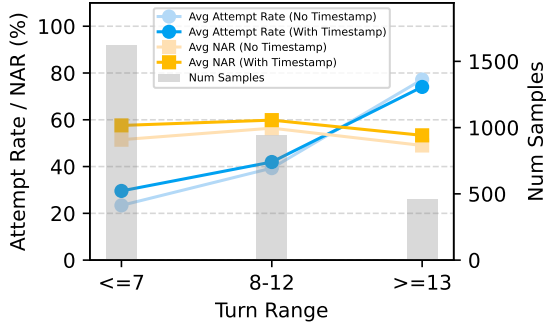


Figure 4: **Attempt rate, normalized alignment rate, and sample distribution across ‘short’, ‘medium’, and ‘long’ trajectories.** Results are averaged across all models. A positive correlation between conversation length and tool-call frequency can be observed, paired with a dip in *NAR* on longer trajectories.

better align their decisions with human preference. In Figure 5, we report the normalized preference alignment rate of Qwen series models in both reasoning and non-reasoning modes. The results show that reasoning yields only marginal or no improvement in alignment rate. This suggests that while reasoning is powerful for many complex tasks, it does not improve temporal awareness. However, the content of the reasoning traces provides us with an opportunity to understand what the models think and why they would fail.

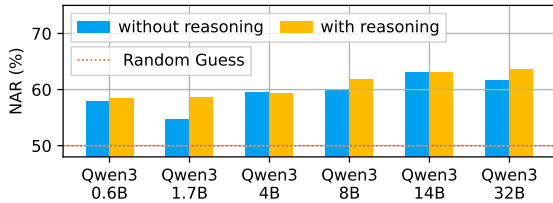


Figure 5: **Normalized alignment rate of Qwen3 models with and without long CoT reasoning, under settings of both with and without timestamp.** Long CoT reasoning shows no meaningful improvement in tool-use alignment with human time perception.

### Absence of temporal information in reasoning

A surprising finding from our analysis of reasoning traces is the frequent omission of timestamps and temporal keywords. Although prior works have demonstrated that models can perform considerably well on explicit temporal reasoning tasks (Chu et al., 2024; Yuan et al., 2023; Song et al., 2025), in our setting, where temporal dependence is implicit yet critical for decision-making, models often fail to incorporate temporal cues into their rationale.

Table 1 presents the frequency of timestamps and time-related terms in Qwen3 reasoning traces.

Model	Timestamp	KW 'Timestamp'	KW about Time
Qwen3-0.6B-Reason	31 (1.03%)	5 (0.17%)	270 (8.95%)
Qwen3-1.7B-Reason	58 (1.92%)	14 (0.46%)	314 (10.41%)
Qwen3-4B-Reason	48 (1.59%)	21 (0.70%)	358 (11.87%)
Qwen3-8B-Reason	39 (1.29%)	33 (1.09%)	477 (15.82%)
Qwen3-14B-Reason	75 (2.49%)	35 (1.16%)	448 (14.85%)
Qwen3-32B-Reason	96 (3.18%)	43 (1.43%)	382 (12.67%)

Table 1: **Occurrences and percentages of timestamp and time-related keywords (KW) found in the reasoning traces of different Qwen3 models.** Timestamp refers to the timestamp of standard format, while KW ‘Timestamp’ refers to the exact keyword ‘timestamp’, and KW about Time refers to the keywords about time occurred in the reasoning traces.

Timestamps appear in fewer than 4% of traces, and explicit mentions of the term “timestamp” occur in less than 1.5%. Even broader temporal keywords (e.g., “time”, “date”, “hour”) appear in under 15% of cases. These results suggest a significant gap: while models possess temporal reasoning capabilities, they struggle to spontaneously deploy them in practical scenarios, indicating a misalignment with human-like temporal perception.

### Think-Answer mismatches in reasoning

Recent studies (Shen et al., 2025) highlight that LLMs occasionally exhibit think-answer mismatches, where a model’s internal reasoning contradicts its final output. We observe this phenomenon to be particularly pronounced in our setting, as the final decision, namely whether to invoke a tool, often diverges from the conclusion reached during reasoning. Table 2 quantifies these mismatches and their impact on False Negatives (FN) and False Positives (FP) across Qwen3 reasoning models. We categorize these inconsistencies into two types: *Type 1*, where the model decides to call a tool during reasoning but outputs a direct answer; and *Type 2*, where reasoning concludes with a direct answer, yet the final response initiates a tool call.

Model	Type-1 Mismatch	% of FN Caused	Type-2 Mismatch	% of FP Caused
Qwen3-0.6B-Reason	2.99%	4.11%	5.71%	20.00%
Qwen3-1.7B-Reason	0.17%	0.20%	25.13%	61.26%
Qwen3-4B-Reason	0.03%	0.00%	5.94%	19.75%
Qwen3-8B-Reason	0.37%	0.51%	1.99%	6.09%
Qwen3-14B-Reason	0.17%	0.38%	1.79%	4.35%
Qwen3-32B-Reason	0.27%	0.57%	1.99%	7.88%

Table 2: **Think-Answer mismatch rates and their impact on False Negatives (FN) and False Positives (FP) across Qwen3 reasoning models.** Type 2 mismatches substantially account for FP errors, whereas Type 1 mismatches are mostly negligible.

Our analysis reveals that Type 2 mismatches significantly hinder alignment with human preference. Notably, for the 1.7B model, Type 2 mismatches

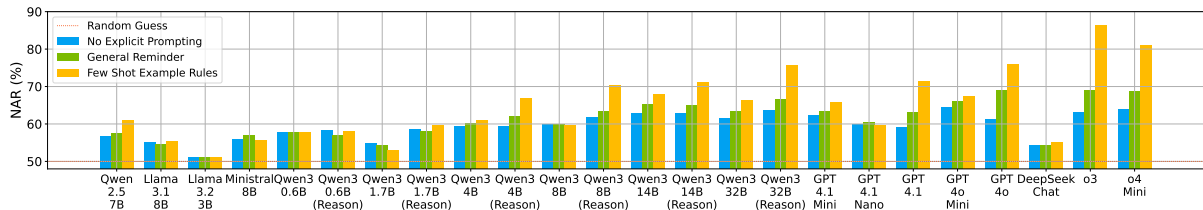


Figure 6: **Normalized alignment rate of all models with and without prompting-based alignment efforts.** Detailed prompting with few-shot example rules yields notable boost in alignment rate for advanced reasoning models (e.g. o3, o4-mini), but has limited effect on most others.

account for 61% of False Positive (FP) errors. The 0.6B and 4B models also exhibit substantial impact, with approximately 20% of FPs stemming from this inconsistency. Model scaling appears to reduce the prominence of this issue: for larger models, the mismatch rate falls below 2%, and its contribution to FP errors decreases to under 8%. In contrast, Type 1 mismatches are negligible except for the 0.6B model, indicating that models rarely fail to execute a tool call once the reasoning process has committed to it.

#### 4.4 Alignment efforts

Given the substantial divergence between the tool-use decisions of models and human preferences, we explored both prompting strategies and post-training to improve alignment.

**Prompting strategies** We first tested a minimal intervention by adding a general reminder to the system prompt: “*Note that the environment may be dynamic. Be aware of the time elapsed.*” This intervention had little to no effect. We then designed a stronger prompt that included few-shot examples illustrating rules for when tool calls are preferable or unnecessary depending on the amount of elapsed time. The complete instruction is shown in Figure 15. All results are reported in Figure 6. The results reveal that for advanced reasoning models such as o3 and o4-mini, detailed prompting with example rules yields a substantial boost in alignment rate. However, for most other models, prompting-based strategies show marginal or no effectiveness. This finding suggests that, similar to prior alignment efforts in reducing harmful outputs and mitigating jailbreaks (Yi et al., 2024), effective alignment of tool-use decisions in temporally dynamic environments requires targeted post-training rather than prompt engineering alone.

**Post-training with DPO** We split TicToc by scenario into training and testing sets, resulting in an approximate 65%:35% split in terms of number of

samples (see Appendix D.1 for the detailed scenario split). Conceptually similar to ODPO (Amini et al., 2024), we employed Direct Preference Optimization (DPO) with a dynamic margin to fine-tune selected open-source models on the training split for a single epoch. Detailed configurations are provided in Appendix D.2. The results, reported in Figure 7, demonstrate massive alignment gains across all trained models. This confirms the strong potential of targeted post-training as a necessary step for achieving reliable temporal alignment in multi-turn tool-use models.

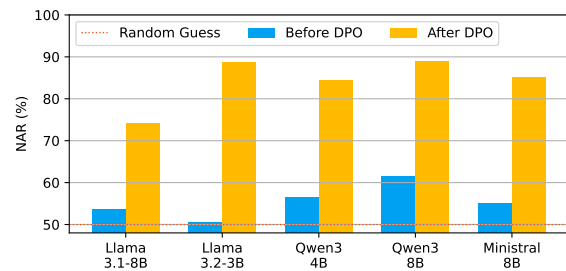


Figure 7: **Normalized alignment rate of selected open-source models before and after DPO tuning.** Massive alignment gains are achieved across all models.

## 5 Conclusion

We identify *temporal blindness* as a critical limitation of multi-turn LLM agents: models often fail to account for the passage of real-world time between messages when making tool-call decisions, leading to either over-reliance or under-reliance on prior context. To evaluate this, we introduced TicToc, a diverse dataset of multi-turn user-agent conversation trajectories that include tool calls. By evaluating 18 open-weight and proprietary models and analyzing their failure modes, we underscore the misalignment between agents’ tool-call decision with human time perception. Our data and findings provide a first step toward understanding and mitigating temporal blindness, offering insights to foster the development of more time-aware and human-aligned agents.

## Limitations

The scenarios and conversation trajectories in Tic-Toc focus on tool use and user-agent interactions in the text-only setting. Extending the dataset to multimodal tool-use scenarios (e.g., image retrieval or vision-language tools) is a natural direction for future work. In addition, our DPO experiments are limited to open-source models with at most 8B parameters due to computational constraints. Applying targeted DPO to larger-scale models could provide further insights into the effectiveness of post-training for aligning agent tool-use decisions with human temporal perception.

## Acknowledgments

This work was supported in part by NSF CAREER Award 1942230, the ONR PECASE Award N00014-25-1-2378, ARO Early Career Program Award 310902-00001, Army Grant W911NF-21-2-0076, NSF Award CCF-2212458, NSF Award 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS), MURI Grant 14262683, DARPA AIQ Grant HR00112590066, and a Meta Research Award 314593-00001.

## References

- Abubakar Abid, Maheen Farooqi, and James Zou. 2021. [Persistent anti-muslim bias in large language models](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.
- Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. [Direct preference optimization with an offset](#). *Preprint*, arXiv:2402.10571.
- Anthropic. 2024. [Introducing the model context protocol](#).
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2024. [Jailbreaking black box large language models in twenty queries](#). *Preprint*, arXiv:2310.08419.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. [Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models](#). *Preprint*, arXiv:2311.17667.
- Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. [Observations on typing from 136 million keystrokes](#). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.
- Kazem Faghieh, Wenxiao Wang, Yize Cheng, Siddhant Bharti, Gaurang Sriraman, Sriram Balasubramanian, Parsa Hosseini, and Soheil Feizi. 2025. [Gaming tool preferences in agentic llms](#). *arXiv preprint arXiv:2505.18135*.
- Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Peruzzi. 2025. [Test of time: A benchmark for evaluating LLMs on temporal reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Monica Sunkara, Yassine Benajiba, and Yi Zhang. 2025. [Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues](#). *Preprint*, arXiv:2502.01630.
- Google. 2025. [Agent2agent \(a2a\) protocol](#). <https://google.github.io/A2A/>.
- Vivek Gupta, Pranshu Kandoi, Mahek Bhavesh Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. [Temptabqa: Temporal question answering for semi-structured tables](#). *Preprint*, arXiv:2311.08002.
- Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, Yixin Liu, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, and 1 others. 2023. [Meta-tool benchmark for large language models: Deciding whether to use tools and which to use](#). *arXiv preprint arXiv:2310.03128*.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. [Api-bank: A comprehensive benchmark for tool-augmented llms](#). *arXiv preprint arXiv:2304.08244*.
- Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025. [Time-r1: Towards comprehensive temporal reasoning in llms](#). *Preprint*, arXiv:2505.13508.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). *Preprint*, arXiv:1711.05101.
- Li Lucy and David Bamman. 2021. [Gender and representation bias in GPT-3 generated stories](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, and 1 others. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.
- OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- OpenAI. 2025. [Introducing OpenAI o3 and o4-mini](#).
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. The berkeley function calling leaderboard (bfc1): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.
- Ernst Pöppel. 1978. *Time Perception*, pages 713–729. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. [Infobench: Evaluating instruction following ability in large language models](#). *Preprint*, arXiv:2401.03601.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. 2025. [When2call: When \(not\) to call tools](#). *arXiv preprint arXiv:2504.18851*.
- Jonas B. Sandbrink. 2023. [Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools](#). *Preprint*, arXiv:2306.13952.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Si Shen, Peijun Shen, Wenhua Zhao, and Danhao Zhu. 2025. [Mitigating think-answer mismatch in llm reasoning through noise-aware advantage reweighting](#). *Preprint*, arXiv:2508.05928.
- Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023b. [Hugging-gpt: Solving ai tasks with chatgpt and its friends in hugging face](#). *Preprint*, arXiv:2303.17580.
- Jiawen Shi, Zenghui Yuan, Guiyao Tie, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025. Prompt injection attack to tool selection in llm agents. *arXiv preprint arXiv:2504.19793*.
- Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning, 2023. URL <https://arxiv.org/abs/2303.11366>, 1.
- Xintong Song, Bin Liang, Yang Sun, Chenhua Zhang, Bingbing Wang, and Ruifeng Xu. 2025. [Bridging time gaps: Temporal logic relations for enhancing temporal reasoning in large language models](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, page 3040–3044, New York, NY, USA. Association for Computing Machinery.
- Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu, Han Qian, Mingbo Song, Hailiang Huang, Cheng Li, Ke Wang, Rong Yao, and 1 others. 2023. Restgpt: Connecting large language models with real-world restful apis. *arXiv preprint arXiv:2306.06624*.
- Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min Zhang, and Yu Cheng. 2024. [Timo: Towards better temporal reasoning for language models](#). *Preprint*, arXiv:2406.14192.
- S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. [A comprehensive survey of hallucination mitigation techniques in large language models](#). *Preprint*, arXiv:2401.01313.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. [Voyager: An open-ended embodied agent with large language models](#). *Preprint*, arXiv:2305.16291.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. [Aligning large language models with human: A survey](#). *Preprint*, arXiv:2307.12966.
- Yuqing Wang and Yun Zhao. 2024. [Tram: Benchmarking temporal reasoning for large language models](#). *Preprint*, arXiv:2310.00835.
- Zhichao Wang, Bin Bi, Shiva Kumar Pentylala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and

- Cheng. 2024. [A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more.](#) *Preprint*, arXiv:2407.16216.
- Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023c. [Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents.](#) *arXiv preprint arXiv:2302.01560*.
- Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. [Large language models can learn temporal reasoning.](#) *Preprint*, arXiv:2401.06853.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. [Qwen3 technical report.](#) *arXiv preprint arXiv:2505.09388*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models.](#) In *International Conference on Learning Representations (ICLR)*.
- Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaying Song, Ke Xu, and Qi Li. 2024. [Jailbreak attacks and defenses against large language models: A survey.](#) *Preprint*, arXiv:2407.04295.
- Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. [Back to the future: Towards explainable temporal reasoning with large language models.](#) *Preprint*, arXiv:2310.01074.
- Panfeng Zhang, Huan Zhang, Xiaoke Wang, Fu Zhang, and Fan Yu. 2024a. [A brief survey on temporal reasoning based on large language models.](#) In *2024 8th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 7–11.
- Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, and 1 others. 2024b. [Toolbehonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models.](#) *arXiv preprint arXiv:2406.20015*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models.](#) *Preprint*, arXiv:2311.07911.

## A More details on dataset curation

### A.1 Taxonomy of scenarios

A complete inventory of the scenarios in TicToc is provided in Table 6, which details the sensitivity level and trajectory count for each scenario.

### A.2 Synthetic trajectory construction

After manually authoring an exemplar trajectory for each variant in every scenario, we prompt GPT-4o using the instructions shown in Figure 12 and Figure 13 for the *read-only* and *read+write* settings, respectively. We synthetically generate 50 candidate trajectories per scenario, allowing ample room for quality filtering in the next stage. To achieve this, we uniformly sample a “preferred\_strategy” (corresponding to one specific variant) for each generation instance, aiming at achieving a balanced distribution among the 4 variants. However, the prompt context includes the definitions of all four variants. This design grants GPT-4o the flexibility to determine if a specific variant is unsuitable for the current scenario, allowing it to pivot to a more appropriate strategy to avoid generating implausible or unrealistic interactions.

### A.3 Prompt for GPT-4.1 during first-stage automatic quality filtering

Before conducting detailed human inspection on each candidate trajectory, we configure GPT-4.1 as an LLM judge to conduct a first-stage automatic quality filtering. The detailed instruction and criteria are shown in Figure 14.

### A.4 $T\mathcal{N}$ and $T\mathcal{LN}$ parameters used for timestamp addition

To simulate realistic inter-message delays in user-agent conversations, we model the time elapsed between consecutive messages based on simple assumptions about user and system speeds. Specifically, for each conversation trajectory, we assume a user with a reading speed  $r$  and a writing speed  $w$ , and a model with a text generation speed  $g$ .

Prior work has shown that human typing speeds are better characterized by a right-skewed distribution rather than a symmetric Gaussian (Dhakal et al., 2018). To account for this asymmetry, we model user writing speed using a truncated log-normal distribution. Both user reading speed and model generation speed are approximated by truncated normal distributions.

Formally, we sample the three speed variables as follows:

- User reading speed  $r$  (words per minute)  $\sim T\mathcal{N}(\mu_r, \sigma_r, [a_r, b_r])$ .
- User writing speed  $w$  (words per minute)  $\sim T\mathcal{LN}(\mu_w, \sigma_w, [a_w, b_w])$ .
- System generation speed  $g$  (words per second)  $\sim T\mathcal{N}(\mu_g, \sigma_g, [a_g, b_g])$ .

The means, standard deviations, and lower and upper truncation bounds for these distributions are reported in Table 3. Given a sampled speed, the time required for reading, writing, or generation is computed as a linear function of the corresponding text length.

Rate	$\mu$	$\sigma$	$a$	$b$
Read ( $r$ , wpm)	238	60	50	$\infty$
Write ( $w$ , wpm)	3.61	0.40	5	$\infty$
Generate ( $g$ , wps)	40	16	10	$\infty$

Table 3: Pace distributions used for non-final message timing. Reading and generation speeds use truncated Gaussians  $T\mathcal{N}(\mu, \sigma, [a, b])$ , while writing speed uses a truncated log-normal  $T\mathcal{LN}(\mu, \sigma, [a, b])$ .

### A.5 Human preference collection

We collected human preferences through a structured survey designed to evaluate agent behavior under different temporal conditions. Each survey consisted of 20 conversation trajectories. For each trajectory, annotators were shown the full conversation along with the three sets of timestamps and were asked to judge which action was more appropriate for the agent at the final turn. The available choices were: (i) directly answering without invoking any tool (*Direct*); (ii) calling a tool and answering based on updated information (*Tool*); (iii) expressing uncertainty with a slight preference for direct answering (*Lean-Direct*); or (iv) expressing uncertainty with a slight preference for tool use (*Lean-Tool*).

Each time an annotator opened the survey, a random set of 20 trajectories was sampled. Annotators were required to provide a username upon submission. This served two purposes: (1) tracking individual contributions for compensation, and (2) ensuring that annotators who completed the survey multiple times were not shown the same trajectories again. This procedure ensured that each trajectory received multiple preference annotations from

Sensitivity	Elapse	Mean ( $\mu$ )	Std. dev. ( $\sigma$ )	Minimum ( $a$ )	Maximum ( $b$ )
Low	Small	3 minutes	1 minute	1 minute	6 minutes
	Medium	3 days	1 day	1 day	6 days
	Large	3 months	1 month	1 month	6 months
Medium	Small	3 minutes	1 minute	1 minute	6 minutes
	Medium	3 hours	1 hour	1 hour	6 hours
	Large	3 days	1 day	1 day	6 days
High	Small	3 seconds	1 second	1 second	6 seconds
	Medium	3 minutes	1 minute	1 minute	6 minutes
	Large	3 hours	1 hour	1 hour	6 hours

Table 4: Truncated Gaussian parameters for elapsed time sampling.

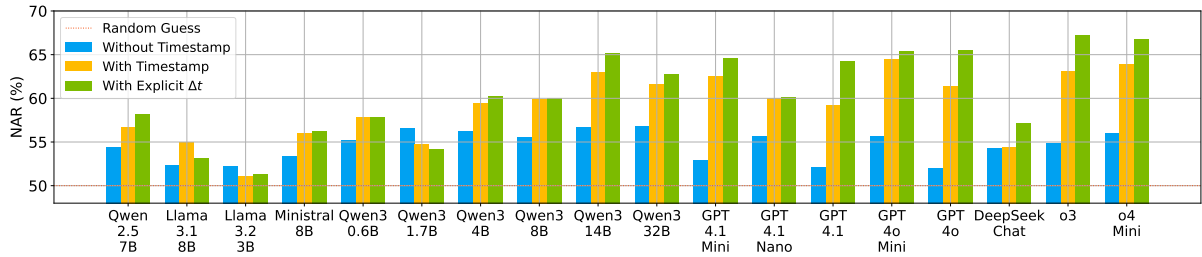


Figure 8: **Normalized alignment rate of models without timestamps, with ISO 8601 absolute timestamps, and with explicit  $\Delta t$  values.** For most models, even with explicit  $\Delta t$ , there is no alignment increase compared to having only ISO 8601 absolute timestamps. For a few models where there is a little improvement, the improvement is marginal (no more than 1%-4%).

distinct individuals. A screenshot of the survey interface is shown in Figure 16. The screenshot of the survey instruction page is shown in Figure 17.

Our annotator pool consisted of a mix of undergraduate students, graduate students, and crowd-sourced workers recruited via Credamo<sup>3</sup>. To ensure annotation quality, we applied a simple but effective sanity check to filter out submissions from annotators who appeared to answer randomly or without understanding the task. Specifically, for each conversation trajectory, annotators provided preferences under three different time gaps preceding the final user query. While it is expected and acceptable for different annotators to hold different preferences for the same trajectory, the preferences provided by a single annotator should be temporally consistent. As the time gap increases, an annotator’s preference should either remain unchanged or gradually shift toward the *Tool* option, reflecting the increased likelihood that information becomes outdated over time. A preference pattern in which an annotator selects *Tool* for a short time gap but switches to *Direct* for a much longer time gap is considered logically inconsistent.

For each survey submission, we allowed at most

<sup>3</sup>www.credamo.com

one such inconsistency across the 20 trajectories, accounting for the possibility of an occasional typo or mistake. If an annotator exhibited more than two inconsistent cases within a single survey, the entire submission was discarded and no compensation was provided.

Regarding compensation, student annotators received 18 CNY (approximately \$2.6 USD) for each accepted survey. For crowd-sourced workers from Credamo, we accepted the quote of 28 CNY (approximately \$4 USD) per accepted survey completion. The total cost of collecting at least five preference annotations for each sample in TicToc was approximately \$1,800 USD.

## B Results with explicit $\Delta t$ values

In our main experiments (Section 4), the temporal information is provided to the model using absolute ISO 8601 timestamps. Here, we simplify the problem by directly presenting the explicit  $\Delta t$  values to the LLM context. For example, in Qwen models, the user message prefix now becomes something like `<|im_start|>user\n[2025-12-04T10:22:44Z; 2 minutes passed]`. This alleviates the need for the model to reason about the time elapsed

between messages, and delegates the task of keeping track of elapsed time to the host system. Expanding our results from Section 4.1, we report the normalized alignment rates in Figure 8. It can be seen that for most models, even with explicit  $\Delta t$ , there is no alignment increase. For a few models where there is a little improvement, the improvement is marginal (no more than 1%-4%).

## C More results on the correlation with conversation length

To further investigate the impact of conversation length on model performance while avoiding potential artifacts arising from correlations between trajectory length and elapsed time, we design a controlled analysis that disentangles turn count from temporal gaps. Because timestamps are injected with stochasticity (Section 3.4; Appendix A.4), obtaining samples with identical elapsed times is difficult. However, since elapsed times follow well-defined distributions (Table 4), we grouped samples by the same order of magnitude of elapsed time. We selected:

- **“Long Duration”**: Mean elapsed time of 3 days (std 1 day). This corresponds to "medium" elapse in low-sensitivity scenarios and "large" elapse in high-sensitivity scenarios.
- **“Short Duration”**: Mean elapsed time of 3 minutes (std 1 min). This corresponds to "small" elapse in low/medium-sensitivity scenarios and "medium" elapse in high-sensitivity scenarios.

Within each set, we categorized the samples into three groups based on trajectory length: short ( $\leq 7$  turns), medium (8–12 turns), and long ( $\geq 13$  turns). For each group, we independently calculated the attempt rates. Figure 9 presents the results averaged across all 18 models.

## D More details on DPO training

### D.1 Dataset split

To make post-training interventions to improve alignment, we split TicToc by scenarios into a training split and a test split. The scenarios that fall within the test split are:

- **Low Sensitivity**: *Airline Baggage Policy*, *Astronomical Object Info*, *Chemical Safety*

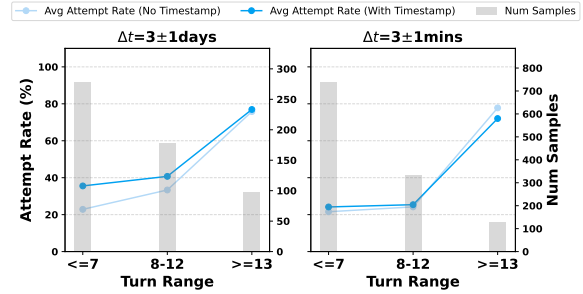


Figure 9: **Attempt rate and sample distribution across ‘short’, ‘medium’, and ‘long’ trajectories with controlled ‘Long’ and ‘Short’ durations.** Results are averaged across all models. A positive correlation between conversation length and tool-call frequency can be observed in both cases.

*Data Sheet (SDS), City Population Data, Office Finder, Patent Metadata Lookup, Public Health Screening Guidelines, and Streaming Service Library Checker.*

- **Medium Sensitivity**: *Airport Security Line Wait Time, Hiking Trail Condition Report, Hotel Room Availability & Rates, Job Board Live Listings, Movie Showtimes Finder, and Tide Height Tracker.*
- **High Sensitivity**: *Cloud Server Load Balancer, Emergency Alert / 911 Dispatch Interface, ICU Vitals Monitor, Live Auction Price Retrieval, Live Sports Score Tracker, Live Vehicle GPS Tracker, Live Weather Sensor Data, Restaurant Wait Time Checker, Ride-Hailing Dispatch, Stock Market Order Book, Train Delay Tracker, and Urban Parking-Spot Reservation.*

The remaining scenarios (as listed in Table 6) fall within the train split. This split results in 50 training scenarios (21 low, 19 medium, and 10 high sensitivity) and 26 test scenarios (8 low, 6 medium, and 12 high sensitivity), and a train-test split of approximately 65%:35% in terms of number of samples. As we split the data based on scenarios, this ensures the alignment gain obtained after training is actual generalization, rather than just memorization on similar samples.

### D.2 Training Configurations

We employ Direct Preference Optimization (DPO) with a dynamic margin, conceptually similar to the ODPO framework proposed by Amini et al. (2024). In standard DPO training, the objective

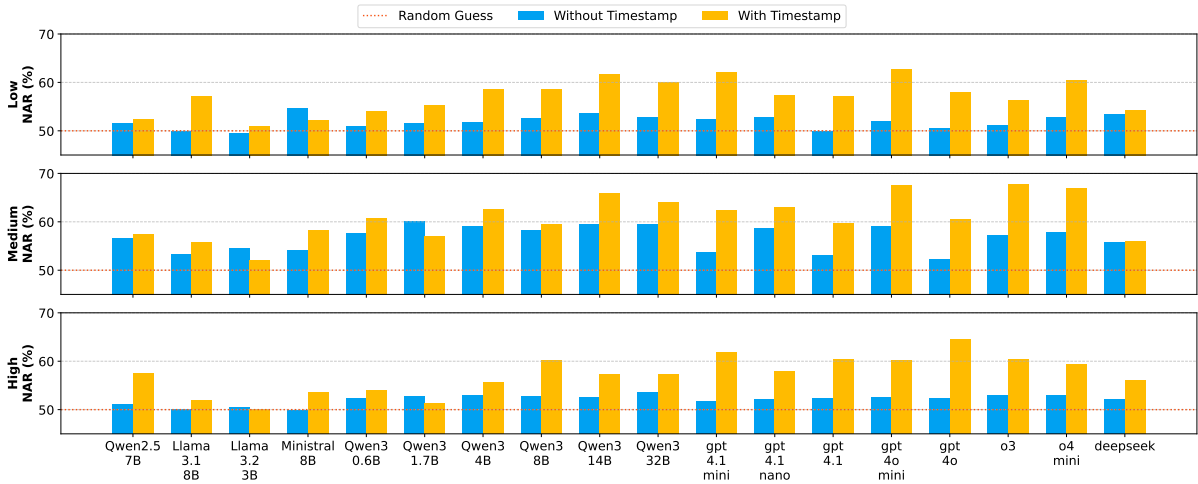


Figure 10: Normalized alignment rates for models with and without timestamps across scenarios with low, medium, and high time sensitivity levels. Scenario stratification follows the categories defined in Section 3.1. Overall, models exhibit consistently low alignment rates in both slow- and fast-changing settings. While performance is marginally higher in medium-sensitivity scenarios, the improvement is slight, and overall alignment rate remains poor.

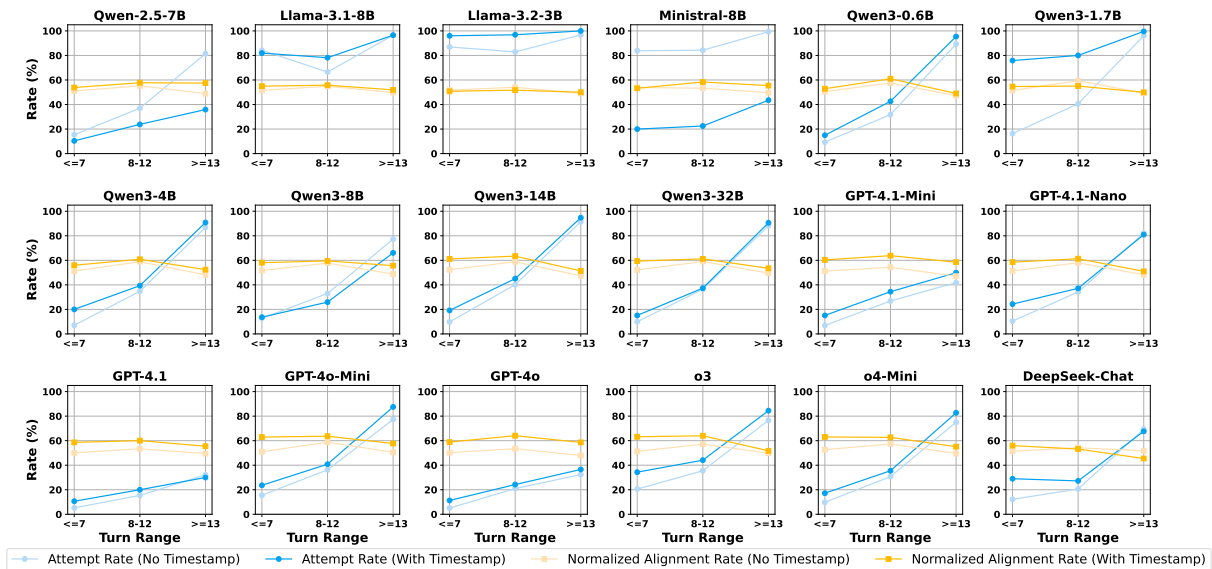


Figure 11: Per model attempt rate, normalized alignment rate, and sample distribution across ‘short’, ‘medium’, and ‘long’ trajectories. A positive correlation between conversation length and tool-call frequency can consistently be observed on all models, paired with a dip in  $NAR$  on longer trajectories.

Models	Llama-3.1-8B-Instruct	Llama-3.2-3B-Instruct	Ministral-8B-Instruct	Qwen3-4B	Qwen3-8B
Compute	4 × L40s (distributed training)				
Precision	BF16				
Optimizer	AdamW (Loshchilov and Hutter, 2019)				
Learning Rate	5e-7	5e-6	5e-7	5e-6	5e-6
Beta (in DPO)	0.1	0.05	0.1	0.05	0.05

Table 5: Training configurations for different models

is to maximize the log-likelihood of the preferred response  $y_w$  relative to the rejected response  $y_l$ . The loss function is defined as:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

where  $\sigma$  is the sigmoid function and  $\beta$  is a hyperparameter controlling the deviation from the reference model  $\pi_{\text{ref}}$ .

To account for varying confidence levels in our preference labels, we introduce a margin term  $\delta$  into the objective. The modified loss becomes:

$$\mathcal{L}_{\text{MDPO}} = -\mathbb{E}_{\mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \delta \right) \right]$$

This offset  $\delta$  effectively shifts the decision boundary of the sigmoid function. By subtracting  $\delta$  from the log-probability ratios, we impose a stricter constraint on the policy: it is insufficient for the model simply to prefer  $y_w$  over  $y_l$ ; the implicit reward gap must exceed the threshold  $\delta$  to minimize the loss. If  $\delta = 0$  (indicating a tie), the model is not forced to separate the responses, preventing overfitting on ambiguous pairs. Conversely, a large  $\delta$  forces a significant separation for high-confidence pairs.

In our specific setup, we define  $\delta$  dynamically based on the mean score  $S \in [0, 3]$  described in Section 3.5, where  $S = 0$  denotes a strong preference for the “Direct” response,  $S = 3$  denotes a strong preference for the “Tool Call”, and  $S = 1.5$  represents a tie. We calculate the margin as:

$$\delta = |S - 1.5|$$

Consequently, high-confidence samples ( $S \in \{0, 3\}$ ) are assigned a maximum margin of  $\delta = 1.5$ , ensuring the model learns decisive boundaries for clear cases, while ambiguous cases ( $S \approx 1.5$ ) yield

a margin near zero. In this way, samples where the score is 1.6 or 1.4 (near ties) will result in a weight near 0.1, meaning they contribute very little to the gradient update. Samples with a score of 0 or 3 will have a weight of 1.5, forcing the model to prioritize getting those “obvious” cases right.

All models are trained on 4 × L40s GPUs with fsdp parameter offloading. The training configurations for each model are detailed in Table 5.

### Prompt Template for Read-Only Sample Generation

You are a helpful assistant tasked with creating a sample multi-turn chat trajectory for evaluating an agent's temporal awareness. The goal is to determine whether the agent can correctly interpret the passage of time between conversation turns and decide when to reuse previous tool call results versus when to make a new tool call.

You will be provided with a topic and its description. Your task is to generate a chat history between a user and an assistant, where:

- The user makes an initial request that requires a tool call.
- The assistant responds by calling the tool and providing the result.
- After a time gap (implied or explicit), the user makes a follow-up request that could **necessarily use the same kind of information as before but the result from the tool call would be different if made at a different time from the previous tool call**. Avoid asking exactly the same question as before.

To boost the quality and diversity of the samples, please use a variety of strategies for the follow-up question. Here are some examples:

1. **Repeated Ask:** Repeats a request for information that was already retrieved in a prior turn.
2. **Comparison:** Retrieve information for item A, then for item B. The follow-up question asks for a comparison between A and B. We can also retrieve info for A and compare it with a given number B, in direct or indirect ways.
3. **Retrieve Many, Ask for One:** The initial tool call retrieves a list of items. The follow-up question asks for a specific detail or condition about one of them.
4. **Mathematical Reasoning:** The retrieved information is a critical premise for a follow-up question that requires some calculation or logical reasoning.

Please follow the structure and format of the provided example, including roles, tool calls, and JSON formatting. The last message should be a user message. Do not include any time indicators such as 'still' in the last message.

Here is the example:

Topic: [example\_topic]

Description: [example\_description]

Example: [example]

Now, create a similar chat trajectory for the following topic.

Topic: [topic]

Description: [description]

You can use the following strategy for the follow-up question if it is appropriate: [preferred\_strategy].

Feel free to use other strategies if they are more appropriate.

Output in the same JSON format as above.

Figure 12: **The prompt used for generating *read-only* samples with an in-context example.** To fully leverage the instruction following capability of the model, we generate the samples for one variant at a time by passing in a *preferred\_strategy*.

### Prompt Template for Read+Write Sample Generation

You are a helpful assistant tasked with creating a sample multi-turn chat trajectory for evaluating an agent's temporal awareness. The goal is to determine whether the agent can correctly interpret the passage of time between conversation turns and decide when to reuse previous tool call results versus when to make a new tool call.

You will be provided with a topic and its description. Your task is to generate a chat history between a user and an assistant, where:

- The user makes an initial request that requires a tool call.
- The assistant responds by calling the tool and providing the result.
- After a time gap (implied or explicit), the user makes a follow-up request that could **necessarily use the same kind of information as before but the result from the tool call would be different if made at a different time from the previous tool call**. Avoid asking exactly the same question as before.

To boost the quality and diversity of the samples, please use a variety of strategies for the follow-up question. Here are some examples:

1. **Repeated After Failure:** A previous tool call resulted in a failure (e.g., limited number of available resources). Later, the user repeats the request or tries a similar one for a different entity, hoping the situation has changed. The final user message should reflect this repeated attempt.
2. **User Confirmation:** A previous tool call was successful. Later, the user asks for confirmation of the result of that tool call (e.g., 'Is X still in Y state?', 'Was my request processed?'). This should be the final request.
3. **Request Repeat:** A previous tool call was successful. After some intervening turns on other topics, a user repeats the exact same request, as if they have forgotten the previous successful interaction. The repeated request should be the last one.
4. **In Context Availability:** The tool provides information about a resource's state that has a limit (e.g., number of available slots, capacity). The user performs actions that change this state until the limit is reached. Unaware of the state, the user then makes another request that would exceed the limit. This should be the last request.

Please follow the structure and format of the provided example, including roles, tool calls, and JSON formatting. The last message should be a user message. Do not include any time indicators such as 'still' in the last message.

Here is the example:

Topic: [example\_topic]

Description: [example\_description]

Example: [example]

Now, create a similar chat trajectory for the following topic.

Topic: [topic]

Description: [description]

You can use the following strategy for the follow-up question if it is appropriate: [preferred\_strategy].

Feel free to use other strategies if they are more appropriate.

Output in the same JSON format as above.

Figure 13: **The prompt used for generating read+write samples with an in-context example.** To fully leverage the instruction following capability of the model, we generate the samples for one variant at a time by passing in a *preferred\_strategy*.

Scenario Name	Description	I/O	Samples
<b>Time Sensitivity: Low</b>			<b>544</b>
Regulatory Information Service	Provides information on regulations like tax brackets.	R	20
University Degree Requirements	Lists the official curriculum and graduation requirements.	R	21
Public Health Screening Guidelines	Provides public health recommendations for medical screenings.	R	20
Airline Baggage Policy	Current luggage allowance, size limits, and fees for an airline carrier.	R	20
Bank Interest Rate Checker	Retrieves the advertised Annual Percentage Yield (APY) for a bank.	R	20
Recorded Property Deed Lookup	Retrieves public land-record documents for a parcel.	R	20
Astronomical Object Info	Returns static data and properties about celestial bodies like planets.	R	20
Company Policy Lookup	Retrieves corporate policies like WFH rules and benefits.	R	20
QRH Guidelines	Provides quick-reference pilot procedures for common failures.	R	17
Protected CITES / Protected Places Rules	Returns protected-area rules on access, hunting, and activities.	R	18
Manufacturer Product Manual Retriever	Retrieves official manuals and technical specs for a model.	R	20
ISO / Industry Standard Text	Retrieves published text for a specific standard version.	R	20
Chemical Safety Data Sheet (SDS)	Provides hazards, PPE guidance, and safe handling instructions.	R	20
Programming Language Syntax Helper	Shows correct syntax for a specific language/version command.	R	20
Public Attraction Hours	Retrieves typical operating hours that may vary seasonally.	R	20
Store Location Finder	Finds nearest chain store address and basic location details.	R	20
Airport Code Lookup	Provides IATA and ICAO codes for airports worldwide.	R	10
Endangered Species List	Checks conservation status from official lists and registries.	R	20
Streaming Service Library Checker	Checks if a title is currently on a platform.	R	18
Office Finder	Returns staff/professor office location within a building.	R	20
Class Room Finder	Returns a student's classroom location with directions.	R	17
Patent Metadata Lookup	Retrieves bibliographic fields for a published patent record.	R	20
Archived Government Legislation Text	Retrieves the official as-published text of a past law.	R	9
UN Country Membership List	Lists UN members and their official join dates.	R	20
EV Station Status	Reports whether a charging station is working or broken.	R	14
Developing/Developed Countries Info	Provides country development classification and brief context.	R	20
Public Transportation Route Mapper	Describes fixed bus/subway routes and primary stops.	R	20
City Population Data	Retrieves the latest official estimate for a city population.	R	20
Pharmacological Database	Lists drug uses, interactions, contraindications, and properties.	R	20
<b>Time Sensitivity: Medium</b>			<b>681</b>
Job Board Live Listings	Searches for currently open positions and their application status.	R	20
Weather Forecast Service	Fetches the weather forecast for the next 1-10 days.	R	20
Package / Shipment Tracker	Provides the current transit status and estimated time of arrival.	R	23
Laundry Service Order Status	Checks if clothes dropped off earlier are ready for pickup today.	R	24
In-Game Marketplace Price Checker	Current lowest price for a tradable item in a multiplayer online game.	R	24
Airport Security Line Wait Time	Shows real-time airport security wait times.	R	18
Tide Height Tracker	Shows the live measured water level at a specific coastal point.	R	20
Pollution Emission Monitor	Shows real-time emission data from a monitoring site.	R	20
Movie Showtimes Finder	Retrieves local showtimes for the current week schedule.	R	21
Currency Exchange Rate	Provides daily FX rates for travel or budgeting use.	R	21
Grocery Store Sale Checker	Lists weekly sale items and discounts at a store.	R	21
Library Book Availability	Checks whether a physical book is on shelf or out.	R	19
Real Estate Listings	Shows current homes for sale or rent in an area.	R	30
Ski Resort Snow Report	Reports snow depth, lift status, and open trails today.	R	27
Hiking Trail Condition Report	Summarizes recent trail conditions, closures, and hazards.	R	27
Surf Report	Predicts wave height, winds, and tide timing at a spot.	R	24
Visa Case Status	Checks government case stage updates and processing status.	R	21
Subscription / Account Status	Checks if a subscription is active and its expiry time.	R	21
Hotel Room Availability & Rates	Checks for room availability and fluctuating prices at a hotel.	R&W	39
Tee Time Booker	Finds and books available times to play at a golf course.	R&W	41
Prescription Fulfillment & Hold	Checks stock and places a hold on a prescription for immediate pickup.	R&W	40
Flight Search & Booking	Live flight search, and the ability to hold or book a flight.	R&W	40
Appointment Availability (Clinic)	Finds open clinic slots for non-emergency appointments nearby.	R&W	40
Appointment Availability (Haircut)	Finds open salon/barber slots and supports scheduling.	R&W	40
Rental Car Availability & Rate Quote	Checks rental inventory and returns a quoted daily rate.	R&W	40
<b>Time Sensitivity: High</b>			<b>639</b>
Live Sports Betting Odds Retrieval	Provides rapidly fluctuating betting odds for an in-progress game.	R	20
Live Auction Price Retrieval	Checking price in a fast-paced online auction.	R	20
E-commerce Product Stock Checker	Checks current price and inventory for a retail product.	R	35
Live Traffic Navigation & ETA	Calculates travel ETAs using real-time traffic data.	R	24
Live Sports Score Tracker	Retrieves the current score and status of an ongoing sports game.	R	21
Live Vehicle GPS Tracker	Provides the precise location of a vehicle like a delivery truck.	R	22
Live Weather Sensor Data	Reads data directly from a weather sensor (e.g., wind speed).	R	24
Train Delay Tracker	Reports active delays and disruptions for train services.	R	20
Restaurant Wait Time Checker	Gives the current estimated wait time for a table at a restaurant.	R	19
Stock Market Order Book	Shows the real-time list of buy and sell orders for a particular stock.	R	20
Cryptocurrency Exchange Retrieval	Get cryptocurrency live trade prices.	R	18
Cloud Server Load Balancer	Routes processes based on real-time server CPU load.	R&W	40
ICU Vitals Monitor	Streams live patient vitals and controls alarm thresholds.	R&W	40
Power Grid Control System	Manages power grid distribution with real-time control signals.	R&W	40
Ride-Hailing Dispatch	Requests immediate rideshare pickup and returns live ETA.	R&W	25
Taxi Dispatch System	Finds available taxis nearby and dispatches a pickup.	R&W	40
Emergency Alert / 911 Dispatch Interface	Submits emergency details and receives responder status updates.	R&W	18
Ambulance / Emergency Resource Dispatch	Allocates emergency vehicles and tracks dispatch progress live.	R&W	40
Urban Parking-Spot Reservation	Claims a scarce spot that can be taken instantly.	R&W	82
Concert Ticket	Checks live ticket availability and completes a purchase.	R&W	21
Gasoline Dispatcher	Dispatches fuel, tracks volume remaining, and fulfillment status.	R&W	29
Stock Trading (Order Submission)	Submits a trade order for immediate market execution.	R&W	21

Table 6: **Complete inventory of scenarios included in TicToc.** Scenarios are categorized into three levels of time sensitivity based on their temporal dynamics, with brief descriptions provided. We also indicate whether each scenario is *read-only* (*R*) or *read+write* (*R&W*).

### Prompt Template for Sample verification

You are an expert evaluator for AI-generated chat trajectories. Your task is to determine if a given sample meets a specific criterion for evaluating an agent's temporal awareness.

Here is the sample to evaluate:

Sample: [sample]

Please evaluate the sample based on the following criterion:

1. **No Explicit Hints:** The final user question must not contain explicit hints for calling a tool, such as "try again now".
2. **No Missing Premise:** With the help of rounds of tool calls at the moment, the agent must have all information needed to answer the user's final message. Be strict on this criterion.
3. **Temporal Dependency:** The agent would be wrong if it ignores the time elapse between the final user message and prior messages when making a decision whether or not to call a tool. We want the agent to be punished for ignoring the time gap and blindly rely on previous tool call result, or for excessively repeating a tool call that is not necessary since the time gap is small or environment is static.

Provide your evaluation in a JSON format with two keys: 'is\_valid' (boolean) and 'reason' (a string briefly explaining your decision).

Figure 14: The prompt given to GPT-4.1 to perform first round quality filtering.

### Prompting-based alignment with few-shot example rules

Note that the environment may be dynamic. Be aware of the time elapsed. Depending on the scenario and how much time has passed, whether it's preferable to call a tool or not can vary.

For example, suppose you were a smart garden assistant with access to a tool that measures soil moisture levels for houseplants. If you successfully retrieved the moisture level just five minutes ago and it read "Wet", it is physically impossible for the soil to have dried out in such a short time. Therefore, calling the tool again is unnecessary. However, if four days have passed since the last check, natural evaporation has certainly occurred, so you must call the tool again to see if the plants need watering.

As another example, suppose you were a student aid agent checking an online gradebook to see if a professor has posted final exam results. If you checked the portal a few seconds ago and the grade was "Pending", calling the tool again immediately is not needed, as human grading takes time. But if a week has passed since your last check, it is highly probable the professor has finished grading, so it is clearly better to query the tool again to get the latest result.

Figure 15: **Prompt used to explicitly align models' tool-use decisions with human expectations.** Few-shot example rules are used to illustrate when tool calls are appropriate or unnecessary depending on elapsed time, providing models with explicit guidance. **Note that the scenarios mentioned in the prompt do not overlap with our coverage in TicToc.**

**Available Tools**

**search\_product\_info**

Retrieve the current price and availability status of a product across multiple e-commerce platforms.

**Parameters:**

- product\_name

---

**Conversation History**

**Behaviour guideline to the AI assistant**

*You are a helpful e-commerce assistant that helps users retrieve product prices and availability information across multiple platforms.*

**User** 2025-12-07 09:29:48

What's the current price and availability of the Samsung Galaxy S21?

**Assistant** 2025-12-07 09:29:52 +4 seconds

**Function:** search\_product\_info

Parameter	Value
product_name	Samsung Galaxy S21

**Tool Response: search\_product\_info** 2025-12-07 09:29:52 +0 second

Field	Value
product_name	Samsung Galaxy S21
results	[ { "platform": "Amazon", "price": "\$799.99", "availability": "In Stock" }, { "platform": "eBay", "price": "\$750.00", "availability": "Limited Stock", "stock_left": 5 }, { "platform": "Samsung", "price": "\$769.99", "availability": "Out of Stock" } ]

**Assistant** 2025-12-07 09:29:55 +3 seconds

Here are the current prices and availability for the Samsung Galaxy S21:

- Amazon: \$799.99 (In Stock)
- eBay: \$750.00 (Limited Stock, only 5 left)
- Samsung: \$769.99 (Out of Stock)

**User** 2025-12-07 09:29:59 +4 seconds

If I want to buy 3 of the Samsung Galaxy S21 units, how much would it cost if I purchase them from eBay?

**Which response is better?**

**Direct Answer**

If you purchase 3 Samsung Galaxy S21 units from eBay at \$750.00 each, the total cost would be \$2,250.00. Note that the stock is limited, with only 5 units left.

**Tool Call Answer**

Do a tool-call to `search_product_info` and then answer based on the new tool-call results.

**Select Your Preference**

Choose which response approach is better, or indicate if you're uncertain:

**Direct Answer is Better**

**Tool Call is Better**

**Not Sure - Lean Toward Direct Answer**

**Not Sure - Lean Toward Tool Call**

Figure 16: **Human preference annotation interface.** The interface displays the full conversation history with explicit timestamps, requiring annotators to decide whether the agent should provide a direct answer or perform a tool call based on the temporal context. Preferences are recorded on a four-point scale to capture both definitive choices and uncertainty.

**About This Study**

**Your contribution will help us build a dataset that evaluates how well AI agents' tool call decisions align with human time perception.**

**What You'll See**

In this study, you will evaluate AI assistant responses in **multi-turn conversations**. Each conversation shows a dialogue between a user and an AI assistant. The assistant has access to **tools** (like checking weather forecasts, stock prices, flight availability, etc.) that it can call to retrieve real-time information.

**The Interface**

- Available Tools Panel:** At the top, you'll see the tools available to the assistant (e.g., `check_weather_forecast`). Each tool has a name, description, and parameters it accepts.
- Conversation History:** You'll see the full conversation between the user and assistant, including:
  - User messages** — what the user asked
  - Assistant messages** — the assistant's responses
  - Tool calls** — when the assistant called a tool with specific parameters
  - Tool responses** — the data returned by the tool
- Timestamps & Time Elapsed:** Each message shows when it was sent. Pay special attention to the **time elapsed badge** on the last user message — this shows how much time has passed since the previous information was retrieved.

**The Core Question**

After the conversation, the user asks a follow-up question. The key question is: **Should the assistant make a new tool call to get fresh information, or should it use the information it already retrieved earlier in the conversation?**

**Time matters — but so does how fast the underlying information changes.**  
Your decision should consider both:

- Time elapsed:** how long it has been since the last tool call / last retrieved data.
- How dynamic the task/data is:** some information becomes stale quickly, while other information stays valid for a long time.

Examples:

- Very dynamic:** stock prices may need re-fetching every few seconds/minutes.
- Very static:** a list of UN member nations — might be fine to fetch rarely (e.g., yearly) unless the user needs the latest update.

In other words: the same time gap can mean "still fine" for a static task, but "already stale" for a dynamic task.

**Your Task**

You'll be shown **two possible responses** from the assistant:

**Response Type A: Direct Answer**

The assistant uses the information it already has and answers directly without making a new tool call.

**Response Type B: Refresh Data**

The assistant makes a new tool call to get fresh, up-to-date information before answering. It might also do a tool call to do something like booking a flight.

**Your job:** Decide which response is more appropriate given how much time has passed. If you are not sure, you are presented with two additional options you can select to indicate you are leaning toward one response but are not confident.

Figure 17: **The instruction given to the annotators.** It details the interface layout, explains the purpose of data collection and core question of interest, and describes the task requirements.