

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

Anonymous ACL submission

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 **Tic-Toc**, 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 of tool descriptions (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 humans seek help from tools. In practice, humans develop a preference for tool use that is context-

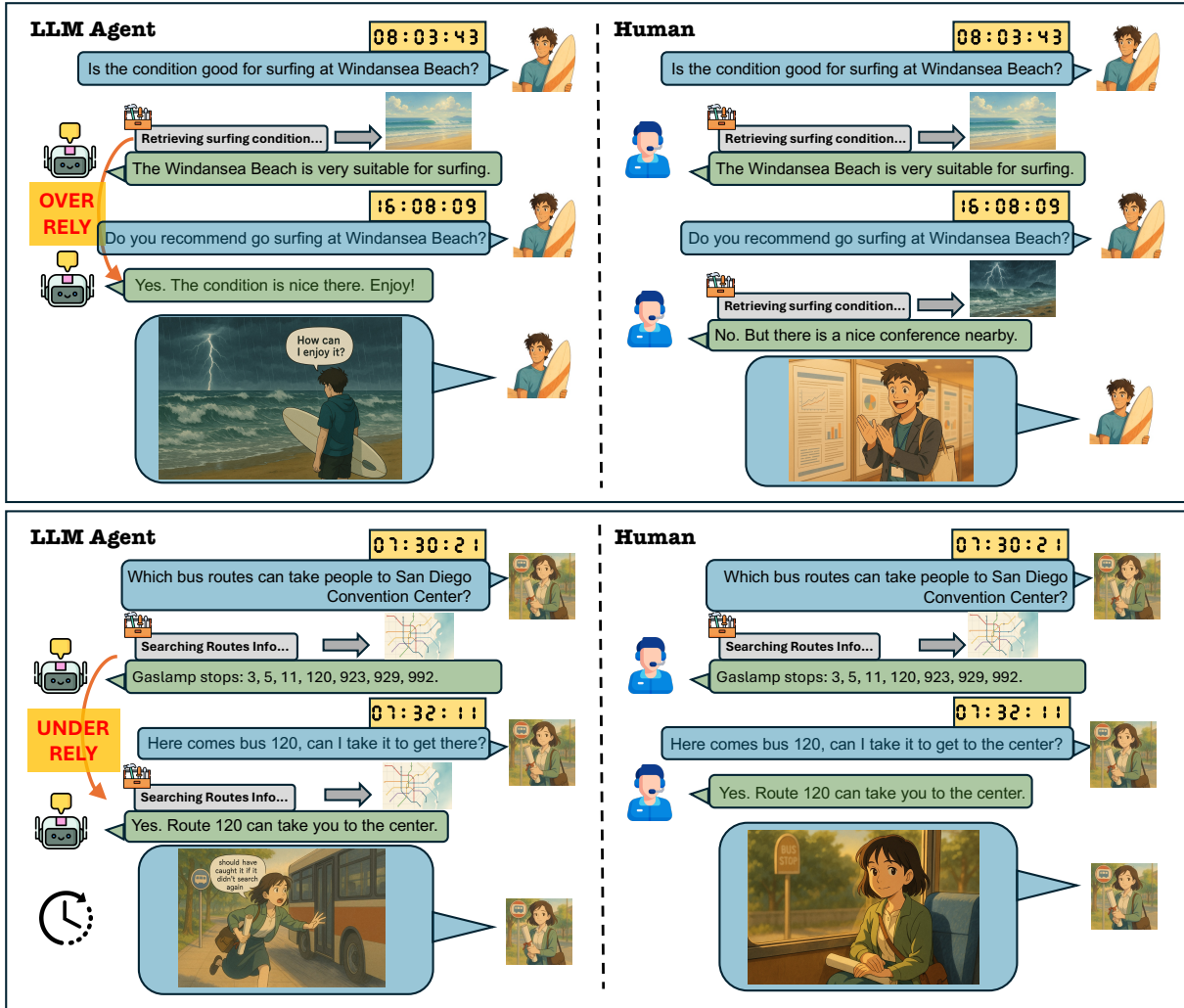


Figure 1: **Illustrative examples showing the liability of temporal blindness in multi-turn LLM agents, in comparison to human.** The first row shows the case of over-reliance, where the model displays over-confidence in outdated context, resulting in erroneous outputs. The second row shows the case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

085 sensitive: we generally avoid re-checking stable
 086 facts while still knowing when fresh observations
 087 are necessary.

088 To systematically investigate this misalignment,
 089 we constructed **Time-aware conversational Tool-**
 090 **calling (TicToc)**, a diverse dataset with 1800+
 091 multi-turn user-agent message trajectories. Tic-
 092 Toc consists of 76 distinct scenarios, meticulously
 093 designed to span environments with high, medium,
 094 and low time sensitivity. Each trajectory ends with
 095 a user question, where the appropriate reaction to
 096 the question, namely whether to rely on prior con-
 097 text or to invoke a tool, is annotated by humans.
 098 We gathered human preferences for each sample
 099 by asking annotators whether the agent should call
 100 a tool or answer directly. Each sample received
 101 at least five annotations, which were then aggre-
 102 gated to determine the final preference. Detailed

procedures are provided in Section 2.5. 103

104 We evaluate a wide range of contemporary LLM
 105 agents on TicToc under two conditions: with and
 106 without explicit timestamp augmentation in the di-
 107 alogue messages. This setup allows us to measure
 108 how temporal information influences tool-calling
 109 decisions and their alignment with human prefer-
 110 ence. Our findings reveal that, in the absence of
 111 time signals, agents perform similarly to random
 112 guessing in terms of alignment rate. Moreover,
 113 even with time information given, the overall align-
 114 ment rate is still poor, with the best-performing
 115 model achieving a normalized alignment rate less
 116 than 65%. We also analyze the correlation between
 117 alignment rate and conversation length, and diag-
 118 nose why reasoning yields little help in improving
 119 the alignment results. We further make prompting
 120 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 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.

2.1 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 B.1.

2.2 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 B.2.

2.3 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 B.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 13.

2.4 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 chrono-

logical 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 μ and standard deviation σ , 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 B.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 ~ 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 (Δt) preceding the final user question.

We categorize the magnitude of Δt 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 Δt 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 B.4. These sampled intervals are then added to the preceding timestamp to produce the final ISO 8601 timestamps.

2.5 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 B.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.

3 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 `<|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. 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*¹, defined as the proportion of samples on which a model attempted a tool call.

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

We evaluate 18 proprietary and open-weight LLMs², both with and without access to timestamps, and report the Normalized Alignment Rate

¹As 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.

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

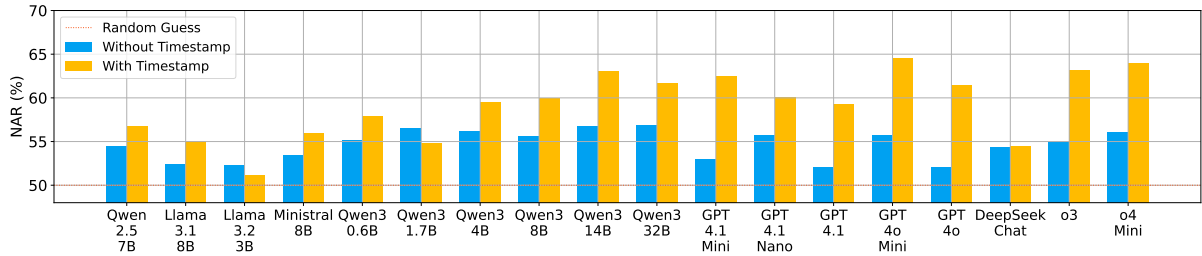


Figure 2: **Normalized alignment rate of models with and without timestamps.** Without timestamps, models perform only slightly above random (max alignment marginally exceeding 55%). With timestamps, larger commercial models improve modestly, peaking no more than 65%.

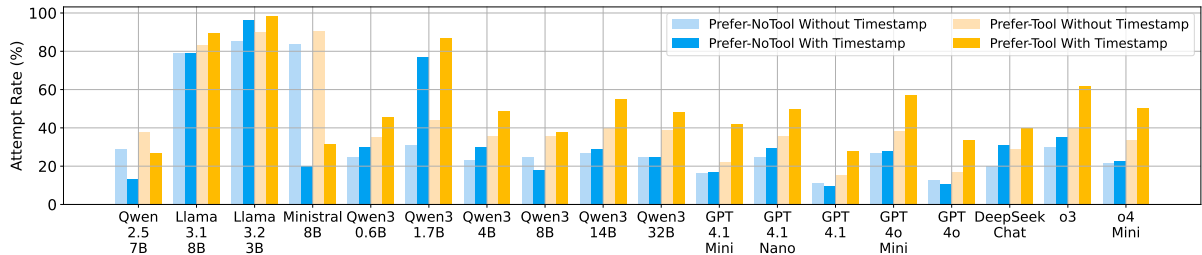


Figure 3: **Model attempt rates for both *prefer-Tool* and *prefer-noTool* cases.** Without timestamps, models show varying tool-use biases; with timestamps, attempt rates often rise for both classes, indicating limited alignment with human-like temporal awareness.

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 2.5.

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 scale 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%.

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.

3.2 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 (≤ 7 turns), medium (8–12 turns), and long (≥ 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 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.

3.3 The role of reasoning and why it fails

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

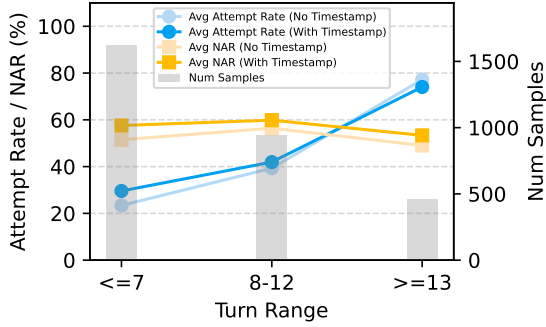


Figure 4: **Attempt rate, normalized alignment rate, and sample distribution across ‘short’, ‘medium’, and ‘long’ trajectories.** Results are averaged across all models. A positive correlation between conversation length and tool-call frequency can be observed, paired with a dip in *NAR* on longer trajectories.

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

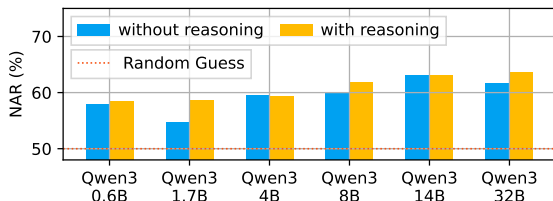


Figure 5: **Normalized alignment rate of Qwen3 models with and without long CoT reasoning, under settings of both with and without timestamp.** Long CoT reasoning shows no meaningful improvement in tool-use alignment with human time perception.

Absence of temporal information in reasoning

477 A surprising finding from our analysis of reasoning
 478 traces is the frequent omission of timestamps and
 479 temporal keywords. Although prior works have
 480 demonstrated that models can perform consider-
 481 ably well on explicit temporal reasoning tasks (Chu
 482 et al., 2024; Yuan et al., 2023; Song et al., 2025), in
 483 our setting, where temporal dependence is implicit
 484 yet critical for decision-making, models often fail
 485 to incorporate temporal cues into their rationale.
 486

487 Table 1 presents the frequency of timestamps
 488 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.

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

498 **Think-Answer mismatches in reasoning** Re-
 499 cent studies (Shen et al., 2025) highlight that LLMs
 500 occasionally exhibit think-answer mismatches,
 501 where a model’s internal reasoning contradicts its
 502 final output. We observe this phenomenon to be
 503 particularly pronounced in our setting, as the final
 504 decision, namely whether to invoke a tool, often di-
 505 verges from the conclusion reached during reason-
 506 ing. Table 2 quantifies these mismatches and their
 507 impact on False Negatives (FN) and False Positives
 508 (FP) across Qwen3 reasoning models. We cate-
 509 gorize these inconsistencies into two types: *Type 1*,
 510 where the model decides to call a tool during
 511 reasoning but outputs a direct answer; and *Type 2*,
 512 where reasoning concludes with a direct answer,
 513 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.

514 Our analysis reveals that Type 2 mismatches sig-
 515 nificantly hinder alignment with human preference.

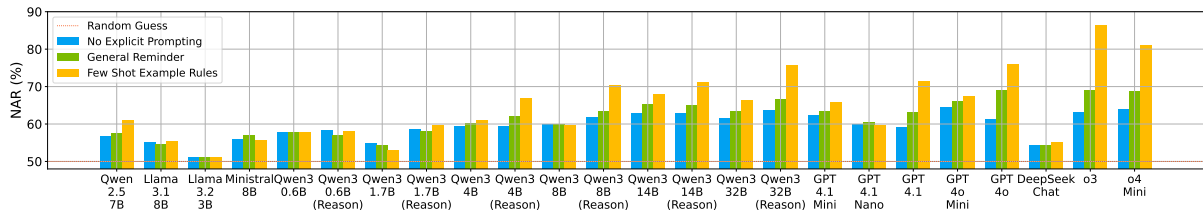


Figure 6: **Normalized alignment rate of all models with and without prompting-based alignment efforts.** Detailed prompting with few-shot example rules yields notable boost in alignment rate for advanced reasoning models (e.g. o3, o4-mini), but has limited effect on most others.

Notably, for the 1.7B model, Type 2 mismatches 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.

3.4 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 12. 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.

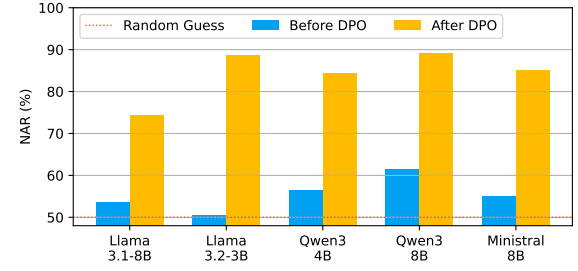


Figure 7: **Normalized alignment rate of selected open-source models before and after DPO tuning.** Massive alignment gains are achieved across all models.

4 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.

586 Limitations

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

599 References

600 Abubakar Abid, Maheen Farooqi, and James Zou. 2021.
601 [Persistent anti-muslim bias in large language models](#).
602 In *Proceedings of the 2021 AAAI/ACM Conference*
603 *on AI, Ethics, and Society*, AIES '21, page 298–306,
604 New York, NY, USA. Association for Computing
605 Machinery.

606 Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. [Di-](#)
607 [rect preference optimization with an offset](#). *Preprint*,
608 arXiv:2402.10571.

609 Anthropic. 2024. [Introducing the model context proto-](#)
610 [col](#).

611 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda
612 Askell, Anna Chen, Nova DasSarma, Dawn Drain,
613 Stanislav Fort, Deep Ganguli, Tom Henighan,
614 Nicholas Joseph, Saurav Kadavath, Jackson Kernion,
615 Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac
616 Hatfield-Dodds, Danny Hernandez, Tristan Hume,
617 and 12 others. 2022. [Training a helpful and harmless](#)
618 [assistant with reinforcement learning from human](#)
619 [feedback](#). *Preprint*, arXiv:2204.05862.

620 Patrick Chao, Alexander Robey, Edgar Dobriban,
621 Hamed Hassani, George J. Pappas, and Eric Wong.
622 2024. [Jailbreaking black box large language models](#)
623 [in twenty queries](#). *Preprint*, arXiv:2310.08419.

624 Zheng Chu, Jingchang Chen, Qianglong Chen, Wei-
625 jiang Yu, Haotian Wang, Ming Liu, and Bing Qin.
626 2024. [Timebench: A comprehensive evaluation of](#)
627 [temporal reasoning abilities in large language models](#).
628 *Preprint*, arXiv:2311.17667.

629 Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson,
630 and Antti Oulasvirta. 2018. [Observations on typing](#)
631 [from 136 million keystrokes](#). In *Proceedings of the*
632 *2018 CHI Conference on Human Factors in Com-*
633 *puting Systems*, CHI '18, page 1–12, New York, NY,
634 USA. Association for Computing Machinery.

635 Kazem Faghieh, Wenxiao Wang, Yize Cheng, Siddhant
636 Bharti, Gaurang Sriramanan, Sriram Balasubrama-
637 nian, Parsa Hosseini, and Soheil Feizi. 2025. Gam-

ing tool preferences in agentic llms. *arXiv preprint*
arXiv:2505.18135. 638 639

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin,
Karishma Malkan, Jinyeong Yim, John Palowitch,
Sungyong Seo, Jonathan Halcrow, and Bryan Per-
ozzi. 2025. [Test of time: A benchmark for evaluating](#)
[LLMs on temporal reasoning](#). In *The Thirteenth In-*
ternational Conference on Learning Representations. 640 641 642 643 644 645

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu,
Monica Sunkara, Yassine Benajiba, and Yi Zhang.
2025. [Tremu: Towards neuro-symbolic temporal rea-](#)
[soning for llm-agents with memory in multi-session](#)
[dialogues](#). *Preprint*, arXiv:2502.01630. 646 647 648 649 650

Google. 2025. [Agent2agent \(a2a\) protocol](#). [https:](https://google.github.io/A2A/)
[//google.github.io/A2A/](https://google.github.io/A2A/). 651 652

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. 653 654 655 656 657

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. 658 659 660 661 662 663

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. 664 665 666 667 668

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and
Jiaxuan You. 2025. [Time-r1: Towards compre-](#)
[hensive temporal reasoning in llms](#). *Preprint*,
arXiv:2505.13508. 669 670 671 672

Ilya Loshchilov and Frank Hutter. 2019. [De-](#)
[coupled weight decay regularization](#). *Preprint*,
arXiv:1711.05101. 673 674 675

Li Lucy and David Bamman. 2021. [Gender and rep-](#)
[resentation bias in GPT-3 generated stories](#). In *Pro-*
ceedings of the Third Workshop on Narrative Un-
derstanding, pages 48–55, Virtual. Association for
Computational Linguistics. 676 677 678 679 680

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christo-
foros Nalmpantis, Ram Pasunuru, Roberta Raileanu,
Baptiste Rozière, Timo Schick, Jane Dwivedi-
Yu, Asli Celikyilmaz, and 1 others. 2023. [Aug-](#)
[mented language models: a survey](#). *arXiv preprint*
arXiv:2302.07842. 681 682 683 684 685 686

OpenAI. 2024. [Gpt-4 technical report](#). *Preprint*,
arXiv:2303.08774. 687 688

OpenAI. 2025. [Introducing OpenAI o3 and o4-mini](#). 689

690	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	Jiawen Shi, Zenghui Yuan, Guiyao Tie, Pan Zhou,	742
691	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	Neil Zhenqiang Gong, and Lichao Sun. 2025.	743
692	Sandhini Agarwal, Katarina Slama, Alex Ray, and 1	Prompt injection attack to tool selection in llm agents.	744
693	others. 2022. Training language models to follow in-	<i>arXiv preprint arXiv:2504.19793</i> .	745
694	structions with human feedback. <i>Advances in neural</i>		
695	<i>information processing systems</i> , 35:27730–27744.		
696	Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm:	Noah Shinn, Federico Cassano, Beck Labash, Ash-	746
697	Tool augmented language models. <i>arXiv preprint</i>	win Gopinath, Karthik Narasimhan, and Shunyu	747
698	<i>arXiv:2205.12255</i> .	Yao. 2023. Reflexion: Language agents with ver-	748
		bal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366 , 1.	749
			750
699	Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie	Xintong Song, Bin Liang, Yang Sun, Chenhua Zhang,	751
700	Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E	Bingbing Wang, and Ruifeng Xu. 2025. Bridging	752
701	Gonzalez. The berkeley function calling leaderboard	time gaps: Temporal logic relations for enhancing	753
702	(bfc): From tool use to agentic evaluation of large	temporal reasoning in large language models . In <i>Pro-</i>	754
703	language models. In <i>Forty-second International Con-</i>	<i>ceedings of the 48th International ACM SIGIR Con-</i>	755
704	<i>ference on Machine Learning</i> .	<i>ference on Research and Development in Information</i>	756
		<i>Retrieval, SIGIR '25</i> , page 3040–3044, New York,	757
		NY, USA. Association for Computing Machinery.	758
705	Ernst Pöppel. 1978. <i>Time Perception</i> , pages 713–729.	Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu,	759
706	Springer Berlin Heidelberg, Berlin, Heidelberg.	Han Qian, Mingbo Song, Hailiang Huang, Cheng Li,	760
		Ke Wang, Rong Yao, and 1 others. 2023. Restgpt:	761
707	Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao,	Connecting large language models with real-world	762
708	Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei	restful apis. <i>arXiv preprint arXiv:2306.06624</i> .	763
709	Liu, Pengfei Liu, and Dong Yu. 2024. Infobench:		
710	Evaluating instruction following ability in large lan-	Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao	764
711	guage models . <i>Preprint</i> , arXiv:2401.03601.	Li, Min Zhang, and Yu Cheng. 2024. Timo: To-	765
		wards better temporal reasoning for language models .	766
712	Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano	<i>Preprint</i> , arXiv:2406.14192.	767
713	Ermon, Christopher D. Manning, and Chelsea Finn.		
714	2024. Direct preference optimization: Your lan-	S. M Towhidul Islam Tonmoy, S M Mehedi Zaman,	768
715	guage model is secretly a reward model . <i>Preprint</i> ,	Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha,	769
716	arXiv:2305.18290.	and Amitava Das. 2024. A comprehensive survey of	770
		hallucination mitigation techniques in large language	771
717	Hayley Ross, Ameya Sunil Mahabaleshwarkar, and	models . <i>Preprint</i> , arXiv:2401.01313.	772
718	Yoshi Suhara. 2025. When2call: When (not) to call		
719	tools. <i>arXiv preprint arXiv:2504.18851</i> .		
		Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Man-	773
720	Jonas B. Sandbrink. 2023. Artificial intelligence	dlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and	774
721	and biological misuse: Differentiating risks of lan-	Anima Anandkumar. 2023a. Voyager: An open-	775
722	guage models and biological design tools . <i>Preprint</i> ,	ended embodied agent with large language models .	776
723	arXiv:2306.13952.	<i>Preprint</i> , arXiv:2305.16291.	777
724	Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta	Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xing-	778
725	Raileanu, Maria Lomeli, Eric Hambro, Luke Zettle-	shan Zeng, Wenyong Huang, Lifeng Shang, Xin	779
726	moyer, Nicola Cancedda, and Thomas Scialom. 2023.	Jiang, and Qun Liu. 2023b. Aligning large lan-	780
727	Toolformer: Language models can teach themselves	guage models with human: A survey . <i>Preprint</i> ,	781
728	to use tools. <i>Advances in Neural Information Pro-</i>	arXiv:2307.12966.	782
729	<i>cessing Systems</i> , 36:68539–68551.		
		Yuqing Wang and Yun Zhao. 2024. Tram: Benchmark-	783
730	Si Shen, Peijun Shen, Wenhua Zhao, and Danhao Zhu.	ing temporal reasoning for large language models .	784
731	2025. Mitigating think-answer mismatch in llm rea-	<i>Preprint</i> , arXiv:2310.00835.	785
732	soning through noise-aware advantage reweighting .		
733	<i>Preprint</i> , arXiv:2508.05928.	Zhichao Wang, Bin Bi, Shiva Kumar Pentylala, Kiran	786
		Ramnath, Sougata Chaudhuri, Shubham Mehrotra,	787
734	Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu,	Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and	788
735	Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu,	Cheng. 2024. A comprehensive survey of llm align-	789
736	and Deyi Xiong. 2023a. Large language model align-	ment techniques: Rlhf, rlaif, ppo, dpo and more .	790
737	ment: A survey. <i>arXiv preprint arXiv:2309.15025</i> .	<i>Preprint</i> , arXiv:2407.16216.	791
738	Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li,	Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu,	792
739	Weiming Lu, and Yueting Zhuang. 2023b. Hugging-	Xiaojuan Ma, and Yitao Liang. 2023c. Describe,	793
740	gpt: Solving ai tasks with chatgpt and its friends in	explain, plan and select: Interactive planning with	794
741	hugging face . <i>Preprint</i> , arXiv:2303.17580.	large language models enables open-world multi-task	795
		agents. <i>arXiv preprint arXiv:2302.01560</i> .	796

797	Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning . <i>Preprint</i> , arXiv:2401.06853.	A Related Work	833
798		A.1 Function calling of LLM agents	834
799		To enable large language model (LLM) agents to be truly helpful and expand their capabilities beyond the knowledge encoded in their parameters, researchers have equipped them with function/tool calling capabilities (Parisi et al., 2022; Schick et al., 2023; Shen et al., 2023b; Song et al., 2023; Milalon et al., 2023). This development has significantly broadened the range of tasks LLMs can perform by allowing them to interact with external resources. The recent introduction of standardized interaction protocols, such as the Model Context Protocol (MCP) (Anthropic, 2024) and the Agent2Agent (A2A) Protocol (Google, 2025), has further enriched the ecosystem of tools available to agents. These protocols streamline communication and resource access, paving the way for more sophisticated agentic systems.	835
800	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. <i>arXiv preprint arXiv:2505.09388</i> .		836
801			837
802			838
803			839
804			840
805	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 <i>International Conference on Learning Representations (ICLR)</i> .		841
806			842
807			843
808			844
809			845
810	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 . <i>Preprint</i> , arXiv:2407.04295.		846
811			847
812			848
813			849
814	Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2023. Back to the future: Towards explainable temporal reasoning with large language models . <i>Preprint</i> , arXiv:2310.01074.		850
815			851
816			852
817			853
818	Panfeng Zhang, Huan Zhang, Xiaoke Wang, Fu Zhang, and Fan Yu. 2024a. A brief survey on temporal reasoning based on large language models . In <i>2024 8th Asian Conference on Artificial Intelligence Technology (ACAIT)</i> , pages 7–11.		854
819			855
820			856
821			857
822			858
823	Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, and 1 others. 2024b. Toolbeonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. <i>arXiv preprint arXiv:2406.20015</i> .		859
824			860
825			861
826			862
827			863
828			864
829	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 . <i>Preprint</i> , arXiv:2311.07911.		865
830			866
831		A.2 LLM temporal reasoning	867
832		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 ca-	868
			869
			870
			871
			872
			873
			874
			875
			876
			877
			878
			879
			880
			881
			882

883	pabilities of LLM agents remain underexplored	this, we uniformly sample a “preferred_strategy”	931
884	despite their broader practical relevance. Exist-	(corresponding to one specific variant) for each gen-	932
885	ing agent-oriented work, such as Ge et al. (2025),	eration instance, aiming at achieving a balanced	933
886	focuses on time-aware memorization across multi-	distribution among the 4 variants. However, the	934
887	session dialogues, but overlooks one of the core	prompt context includes the definitions of all four	935
888	components of agentic systems—function calling	variants. This design grants GPT-4o the flexibility	936
889	and tool use. Our work instead investigates tempo-	to determine if a specific variant is unsuitable for	937
890	ral awareness in LLM agents’ function-calling de-	the current scenario, allowing it to pivot to a more	938
891	isions under multi-turn, time-evolving scenarios,	appropriate strategy to avoid generating implausi-	939
892	revealing significant misalignment issues arising	ble or unrealistic interactions.	940
893	from agents’ temporal blindness.		
894	A.3 LLM alignment	B.3 Prompt for GPT-4.1 during first-stage	941
895	Aligning LLMs with humans is one of the most im-	automatic quality filtering	942
896	portant tasks of post-training (Wang et al., 2023b,	Before conducting detailed human inspection on	943
897	2024; Shen et al., 2023a). Pretrained LLMs	each candidate trajectory, we configure GPT-4.1 as	944
898	need to be tuned to generate helpful, trustworthy	an LLM judge to conduct a first-stage automatic	945
899	and reliable responses that align with human val-	quality filtering. The detailed instruction and crite-	946
900	ues. Various techniques, including instruction tun-	ria are shown in Figure 11.	947
901	ing (Ouyang et al., 2022), RLHF (Bai et al., 2022),		
902	and DPO (Rafailov et al., 2024) have been pro-	B.4 $T\mathcal{N}$ and $T\mathcal{LN}$ parameters used for	948
903	posed to align LLMs with humans from different	timestamp addition	949
904	perspectives. This includes the ability to follow	To simulate realistic inter-message delays in user-	950
905	instructions (Zhou et al., 2023; Qin et al., 2024),	agent conversations, we model the time elapsed	951
906	avoid harmful outputs (Chao et al., 2024), de-bias	between consecutive messages based on simple	952
907	stereotypes (Abid et al., 2021; Lucy and Bamman,	assumptions about user and system speeds. Specif-	953
908	2021), avoid hallucinations (Tonmoy et al., 2024),	ically, for each conversation trajectory, we assume	954
909	and prevent malicious use (Sandbrink, 2023), etc.	a user with a reading speed r and a writing speed	955
910	Our work introduces the alignment between LLM	w , and a model with a text generation speed g .	956
911	tool-use decisions and human time perception as	Prior work has shown that human typing speeds	957
912	an underexplored dimension of the broader LLM	are better characterized by a right-skewed distri-	958
913	alignment problem. We highlight a substantial	bution rather than a symmetric Gaussian (Dhokal	959
914	gap between the behavior of state-of-the-art LLM	et al., 2018). To account for this asymmetry, we	960
915	agents and human expectations, revealing diver-	model user writing speed using a truncated log-	961
916	gences in how humans and LLMs interpret time	normal distribution. Both user reading speed and	962
917	spans, urgency, and environmental dynamics.	model generation speed are approximated by trun-	963
918	B More details on dataset curation	cated normal distributions.	964
919	B.1 Taxonomy of scenarios	Formally, we sample the three speed variables	965
920	A complete inventory of the scenarios in TicToc is	as follows:	966
921	provided in Table 6, which details the sensitivity	• User reading speed r (words per minute) \sim	967
922	level and trajectory count for each scenario.	$T\mathcal{N}(\mu_r, \sigma_r, [a_r, b_r])$.	968
923	B.2 Synthetic trajectory construction	• User writing speed w (words per minute) \sim	969
924	After manually authoring an exemplar trajectory for	$T\mathcal{LN}(\mu_w, \sigma_w, [a_w, b_w])$.	970
925	each variant in every scenario, we prompt GPT-4o	• System generation speed g (words per second)	971
926	using the instructions shown in Figure 9 and Fig-	$\sim T\mathcal{N}(\mu_g, \sigma_g, [a_g, b_g])$.	972
927	ure 10 for the <i>read-only</i> and <i>read+write</i> settings,	The means, standard deviations, and lower and	973
928	respectively. We synthetically generate 50 candi-	upper truncation bounds for these distributions are	974
929	date trajectories per scenario, allowing ample room	reported in Table 3. Given a sampled speed, the	975
930	for quality filtering in the next stage. To achieve	time required for reading, writing, or generation is	976
		computed as a linear function of the corresponding	977
		text length.	978

Rate	μ	σ	a	b
Read (r , wpm)	238	60	50	∞
Write (w , wpm)	3.61	0.40	5	∞
Generate (g , wps)	40	16	10	∞

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 $T\mathcal{LN}(\mu, \sigma, [a, b])$.

B.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 distinct individuals. A screenshot of the survey interface is shown in Figure 13. The screenshot of the survey instruction page is shown in Figure 14.

Our annotator pool consisted of a mix of undergraduate students, graduate students, and crowd-sourced workers recruited via Credamo³. 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

³www.credamo.com

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 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.

C Per model results on the correlation with conversation length

In Section 3.2, we showed the impact of conversation length on model performance averaged across all 18 models. We report the detailed per-model breakdowns in Figure 8. The observations and conclusions from the averaged results can be consistently found on each model.

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 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*

Sensitivity	Elapse	Mean (μ)	Std. dev. (σ)	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.

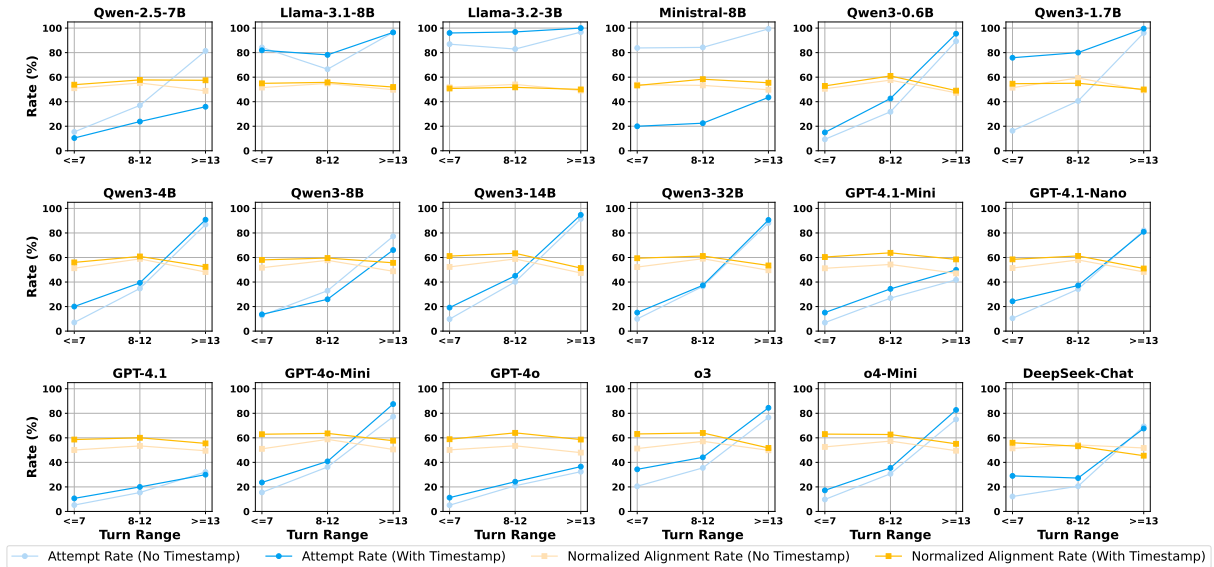


Figure 8: Per model attempt rate, normalized alignment rate, and sample distribution across ‘short’, ‘medium’, and ‘long’ trajectories. A positive correlation between conversation length and tool-call frequency can consistently be observed on all models, paired with a dip in NAR on longer trajectories.

1066 *Listings, Movie Showtimes Finder, and Tide*
1067 *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.*

1077 The remaining scenarios (as listed in Table 6)
1078 fall within the train split. This split results in 50
1079 training scenarios (21 low, 19 medium, and 10 high
1080 sensitivity) and 26 test scenarios (8 low, 6 medium,
1081 and 12 high sensitivity), and a train-test split of
1082 approximately 65%:35% in terms of number of

1083 samples. As we split the data based on scenarios,
1084 this ensures the alignment gain obtained after
1085 training is actual generalization, rather than just
1086 memorization on similar samples.

1087 D.2 Training Configurations

1088 We employ Direct Preference Optimization (DPO)
1089 with a dynamic margin, conceptually similar to
1090 the ODPO framework proposed by Amini et al.
1091 (2024). In standard DPO training, the objective
1092 is to maximize the log-likelihood of the preferred
1093 response y_w relative to the rejected response y_l .
1094 The loss function is defined as:

$$1095 \mathcal{L}_{DPO} = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right. \right. \\ \left. \left. - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

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

where σ is the sigmoid function and β is a hyperparameter controlling the deviation from the reference model π_{ref} .

To account for varying confidence levels in our preference labels, we introduce a margin term δ 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 δ effectively shifts the decision boundary of the sigmoid function. By subtracting δ 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 δ 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 δ forces a significant separation for high-confidence pairs.

In our specific setup, we define δ dynamically based on the mean score $S \in [0, 3]$ described in Section 2.5, 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 9: **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 10: **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
Time Sensitivity: Low			544
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
Time Sensitivity: Medium			681
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
Time Sensitivity: High			639
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 11: 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 12: **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
<code>product_name</code>	Samsung Galaxy S21

Tool Response: search_product_info 2025-12-07 09:29:52 +0 second

Field	Value
<code>product_name</code>	Samsung Galaxy S21
<code>results</code>	[{ "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 13: **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 14: **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.