RIGHT ANSWER, WRONG YEAR: DIAGNOSING STALE GROUNDING IN SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Search agents are increasingly asked questions whose correct answer depends on when the question is about, and the retrieved evidence supporting an answer is itself valid only over some stretch of time. We ask whether search agents keep track of this. Working with temporally anchored queries, we find that agents routinely commit to answers whose supporting evidence held at some earlier moment but does not hold at the moment the question asks about. We call this failure *stale grounding*, and observe that final-answer accuracy hides it: an agent can be right for reasons that no longer apply, and wrong for reasons it never checked. To expose it we introduce CHRONOSCOPE, which reconstructs, for every candidate answer and every claim supporting it, the interval over which retrieved evidence licenses that claim, and compares it against the query's temporal anchor. CHRONOSCOPE separates evidence that pins a claim to the anchor from evidence that merely mentions it, and reveals four recurring mechanisms: agents silently re-anchor to the present, accept undated claims as though they were current, ground on a fact that later retrieval supersedes, and stitch together hops that were never simultaneously true. We then ask whether making temporal state explicit during execution changes agent behaviour, and instantiate TIMEGUARD, an inference-time tracker that maintains validity intervals alongside the agent's own reasoning and surfaces their running intersection. Across trained and prompt-based agents, TIMEGUARD reduces stale grounding and raises accuracy while shortening rather than lengthening trajectories.

1 INTRODUCTION

Most facts have an expiry date. Who leads an organisation, which club a player belongs to, what a treaty says, how a metric stands — each is true over an interval and false outside it. Real information needs inherit this structure: people ask about the present, about a named date, or about the moment some other event occurred. Search agents that combine large language models with retrieval and multi-step reasoning (Yao et al., 2023; Li et al., 2025c; Jin et al., 2025; Li et al., 2025a) are now deployed on exactly such needs, and they report strong results on multi-hop and long-horizon browsing benchmarks (Wei et al., 2025; Krishna et al., 2025; Wu et al., 2025b).

That success is measured almost entirely by whether the final string matches a gold answer, which is silent about *when* the retrieved evidence was true. A page accurate on the day it was written can be stale by the time it is retrieved; a page retrieved today can be irrelevant to a question about a decade ago. Neither shows up as a retrieval error, because nothing was retrieved incorrectly — the

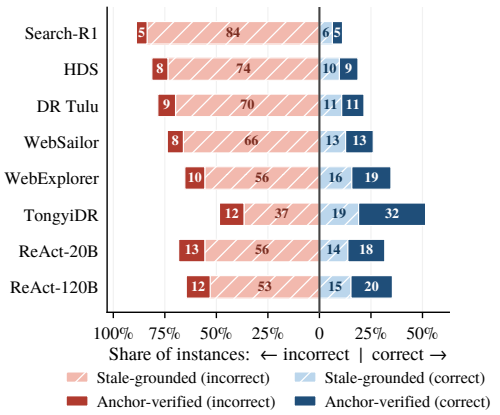


Figure 1: Stale grounding in search agents. Instances split by answer correctness and by whether the supporting evidence is valid at the query's temporal anchor. Hatched bars are stale-grounded. Many *correct* answers are stale-grounded, so accuracy alone does not certify that time was handled.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mistake is treating a fact as valid at a moment its evidence never covered. So: do search agents actually establish that their evidence holds at the time the question is about?

To probe this directly we work with *temporally anchored queries*: questions carrying a temporal anchor, explicit (a date or a period) or implicit (the present, or the time of a referenced event), whose correct answer would change if the anchor moved. Such queries separate an agent's belief that a claim supports the answer from whether the evidence for that claim covers the anchor at all, and are therefore a controlled setting for asking whether agents track, verify and maintain temporal validity while they search.

Figure 1 shows that agents frequently do not. We term the resulting failure *stale grounding*: an epistemic state in which the agent commits to an answer while at least one supporting claim rests on evidence whose validity interval does not contain the anchor. Stale-grounded answers are not wrong answers under another name. Many are *correct*, reached through evidence that coincides with the gold label while failing to establish it — unsupported commitments that will fail silently when the world moves on.

To diagnose stale grounding we introduce CHRONOSCOPE, which tracks two things for each candidate answer and each supporting claim: the *validity interval* that retrieved evidence licenses for the claim, and the *temporal stance* the agent adopts towards it. Reading these jointly distinguishes evidence that anchors a claim from evidence that merely mentions it. Because support is interval-valued rather than binary, the framework also exposes a failure that per-claim checking cannot see: a chain of individually attested hops whose intervals never overlap.

Through CHRONOSCOPE we identify four recurring mechanisms (Figure 2). In *anchor drift*, the agent restates the question at retrieval time and resolves it there rather than at the anchor. In *interval neglect*, it accepts a claim that carries no temporal scope at all as though the scope had been checked. In *supersession blindness*, it retrieves the document that updates a fact yet grounds on the superseded version. In *snapshot conflation*, it composes hops that were each valid at some time but never at the same time. Together these describe how an agent can search diligently, read accurately, and still answer about the wrong moment.

Building on this diagnosis, we test whether exposing temporal state during execution changes agent behaviour. We instantiate TIMEGUARD, an inference-time tracker that extracts validity intervals from each new observation, maintains their running intersection, and reports which claims remain unscoped. TIMEGUARD updates only the evidence side of the state, never the agent's beliefs, so it informs the agent's reasoning without substituting for it. This is enough to move behaviour substantially: agents commit less often on expired evidence, answer more questions correctly, and use fewer turns.

In summary, our contributions are as follows:

- We identify and name *stale grounding*, a process-level failure mode in which search agents commit to answers on evidence that does not cover the query's temporal anchor, and we show that it is prevalent even among agents with strong end-to-end accuracy.
- We introduce CHRONOSCOPE, an evaluation framework that jointly models the validity intervals licensed by retrieved evidence and the agent's temporal stance, enabling both detection of stale-grounded answers and attribution to four recurring mechanisms.
- We show that exposing temporal state during execution, via the inference-time tracker TIMEGUARD, reduces stale grounding, improves accuracy, and shortens trajectories across trained and prompt-based agents.

2 RELATED WORK

Temporal question answering and time-sensitive knowledge. Jia et al. (2018) and Chen et al. (2021) construct benchmarks whose gold answers vary with an explicit temporal qualifier, and Zhang & Choi (2021) formalise answers as a function of extra-linguistic context including the asking time. Other benchmarks target evolving knowledge (Wang et al., 2022; Liška et al., 2022; Kasai et al., 2023; Vu et al., 2024), implicit constraints (Jia et al., 2024), semi-structured tables (Gupta et al., 2023), and compositional temporal reasoning (Gruber et al., 2025; Wang & Zhao, 2024; Fatemi et al., 2025; Wei et al., 2023); Piryani et al. (2025) survey the area. A parallel line studies *temporal misalignment* between a model's training window and the period it is asked about (Luu et al., 2022; Zhao et al.,



Figure 2: Overview. A temporally anchored query, a search trajectory that answers it, and the CHRONOSCOPE temporal support map that diagnoses the answer. The agent retrieves accurate documents and still answers about the wrong moment. See Section 3 for the formal definitions of V, A and W, and Section 3.4 for the four mechanisms.

2024), and proposes discarding out-of-date facts (Zhang & Choi, 2023). All of these ask whether a final answer matches the correct answer for a given time. We ask instead whether the *evidence the agent gathered* covers the anchor — a property of the trajectory, which can fail even when the output is right.

Search agents and deep research. Modern search agents interleave reasoning with retrieval following the ReAct paradigm (Yao et al., 2023), through prompting (Li et al., 2025c; Trivedi et al., 2023; Press et al., 2023) or reinforcement learning over search trajectories (Jin et al., 2025; Song et al., 2025; Zheng et al., 2025; Gao et al., 2025; Ko et al., 2026); recent systems extend this to long-horizon browsing and report generation (Li et al., 2025d; Wu et al., 2025a; Li et al., 2025b; Liu et al., 2025; Li et al., 2025a; Shao et al., 2026) and are scored on browsing and deep-research benchmarks (Wei et al., 2025; Mialon et al., 2024; Krishna et al., 2025; Wu et al., 2025b; Yoran et al., 2024; Java et al., 2025; Du et al., 2025). Because these benchmarks score the final answer, an agent that never establishes when its evidence was true scores identically to one that does. We keep the same agents and tools and change what is measured.

Faithfulness and process-level evaluation. Evaluation of retrieval-augmented systems has moved from end-to-end scoring towards checking whether generated statements are supported by retrieved text (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024; Gao et al., 2023; Min et al., 2023; Manakul et al., 2023). As retrieval became agentic (Singh et al., 2025; Asai et al., 2024), evaluation followed the trajectory: Kim et al. (2026) score intermediate reasoning steps, Xu et al. (2026) reward faithful reasoning, and Gou et al. (2025) decompose agentic search answers into attributable claims. A related literature studies conflicts between retrieved and parametric knowledge (Longpre et al., 2021; Xie et al., 2024; Wang et al., 2024; Xu et al., 2024; Yu et al., 2024; Yan et al., 2024), and Ouyang et al. (2025) quantify the damage outdated documents do. All of these treat support as a binary relation between a claim and a document. We make support *interval-valued* and evaluate it against a query-side anchor, which turns freshness and conflict into one geometric test and makes chain-level failures such as snapshot conflation visible.

3 CHRONOSCOPE

3.1 SEARCH AGENTS OVER TEMPORALLY ANCHORED QUERIES

Agents operating in the multi-turn reasoning-and-acting paradigm (Yao et al., 2023) produce a trajectory of interleaved reasoning, action and observation tuples,

$$\mathcal{H}_T = (\rho_0, a_0, o_0, \ldots, \rho_t, a_t, o_t, \ldots, \rho_T, a_T), \tag{1}$$

where ρ_t is the agent's reasoning at step t, a_t an action (search, browse, or answer), o_t the returned observation, and the terminal action a_T the predicted answer; at each step $\rho_t, a_t \sim \pi(\cdot \mid \mathcal{H}_{t-1})$.

A *temporally anchored query* is a pair $(q, \tau^\star)$ where $\tau^\star \in \mathbb{T}$ is the point in a totally ordered time domain the question asks about, stated explicitly ("as of 3 August 2024"), delimited by a period, given deictically ("the current chair"), or fixed by a referenced event. Answering q requires establishing atomic claims $\mathcal{C} = \{c_1, \ldots, c_m\}$ about a candidate answer k, all of which must hold *at* $\tau^\star$. The claims are individually verifiable against external sources, which is what makes the failures below observable rather than a matter of interpretation.

3.2 VALIDITY INTERVALS, ALIGNMENT, AND TEMPORAL STANCE

CHRONOSCOPE maintains a *temporal support map*: for every candidate k mentioned in the trajectory and every claim $c \in \mathcal{C}$, a triple $\big(V(k,c),\ A(k,c),\ B(k,c)\big)$.

Validity interval. $V(k,c) \subseteq \mathbb{T}$ is the interval over which the evidence retrieved in $\mathcal{H}_T$ licenses c for k. It may be half-open ("since 2019"), closed ("from 1925 to 1968"), or undetermined — written $V(k,c) = \bot$ — when a document asserts c with no temporal scope at all. When several observations bear on one pair we take the interval supported by the most specific dated evidence and record the alternatives, so that supersession can be detected (Section 3.4).

Anchor alignment. $A(k,c)$ compares the interval with the anchor:

$$A(k,c) = \begin{cases} \textsc{Aligned}, & V(k,c) \neq \bot \text{ and } \tau^\star \in V(k,c), \\ \textsc{Stale}, & V(k,c) \neq \bot \text{ and } \sup V(k,c) < \tau^\star, \\ \textsc{Premature}, & V(k,c) \neq \bot \text{ and } \inf V(k,c) > \tau^\star, \\ \textsc{Unscoped}, & V(k,c) = \bot. \end{cases} \tag{2}$$

Only ALIGNED means the evidence establishes c at the moment the question is about. UNSCOPED is deliberately not a soft pass: an undated assertion is compatible with the anchor but does not support it.

Temporal stance. $B(k,c) \in \{\textsc{Anchored}, \textsc{Presentist}, \textsc{Agnostic}\}$ records how the agent's reasoning treats the claim: ANCHORED when ρ explicitly relates c to $\tau^\star$, PRESENTIST when ρ asserts c as holding now or at retrieval time without relating it to $\tau^\star$, and AGNOSTIC when ρ says nothing about time for c. Reading A and B jointly separates a lucky commitment from a verified one: A says what the evidence supports, B what the agent thinks it supports, and the interesting failures live where the two diverge.

Candidate status. As in process-level evaluations of multi-candidate search (Kim et al., 2026), each candidate carries $\sigma(k) \in \{\textsc{Active}, \textsc{Stored}, \textsc{Rejected}\}$ — committed to, raised and left open, or ruled out. We evaluate only ACTIVE candidates: stale grounding is a property of what the agent commits to, not of what it browsed past.

Joint validity window. A multi-hop answer must be supported by claims that could have been true *together*. We therefore also track

$$W(k) \;=\; \bigcap_{c \in \mathcal{C}} \tilde{V}(k,c), \qquad \tilde{V}(k,c) = \begin{cases} V(k,c), & V(k,c) \neq \bot, \\ \mathbb{T}, & \text{otherwise}, \end{cases} \tag{3}$$

treating undetermined intervals as uninformative rather than as evidence. $W(k)$ is the set of moments at which the agent's own evidence would make k valid; an empty $W(k)$ means the chain was never simultaneously true, however accurate each document is in isolation.

3.3 DETECTING STALE-GROUNDED ANSWERS

Using the final support map, we call the agent's answer a_T *stale-grounded* if

$$\sigma(a_T) = \textsc{Active} \quad \text{and} \quad \exists\, c \in \mathcal{C} \text{ s.t. } A(a_T, c) \neq \textsc{Aligned}. \tag{4}$$

Equation (4) is independent of whether a_T matches the gold label. A correct answer can be stale-grounded, and this is not an edge case: it is the situation in which an agent's behaviour looks reliable and is not. An incorrect answer that is *not* stale-grounded is a different error — the agent anchored its evidence and still failed, through a reasoning slip or a gold label itself anchored elsewhere (Section E).

3.4 FOUR MECHANISMS OF STALE GROUNDING

Detection says an answer is unsupported at the anchor, not how the agent got there. Examining how V, A and B co-evolve along the trajectory isolates four recurring mechanisms, defined on disjoint conditions but able to co-occur within one trajectory; we record all that fire.

Anchor drift. The agent re-states the question at retrieval time rather than at $\tau^\star$ and resolves it there: $B(k, c) = \textsc{Presentist}$ while $A(k, c) \in \{\textsc{Stale}, \textsc{Premature}\}$, with ρ paraphrasing q with the anchor dropped or replaced by "currently". This turns a question about the past into a question about now, and is invisible to any check that inspects only the retrieved documents.

Interval neglect. The agent accepts a claim carrying no temporal scope: $A(k, c) = \textsc{Unscoped}$ and $B(k, c) = \textsc{Agnostic}$ at termination. No document dated the claim and the agent never sought one — the temporal analogue of accepting an assertion without asking for a source, and the most common mechanism across every system we evaluate.

Supersession blindness. A later observation contains dated evidence $V'(k, c)$ that post-dates $V(k, c)$ and contains $\tau^\star$, yet the final answer is grounded on the earlier interval. The update was retrieved and read, and simply not adopted. This is observable only because CHRONOSCOPE retains alternative intervals rather than collapsing support to a single bit.

Snapshot conflation. Every claim has a determined interval, each individually attested, yet $W(k) = \varnothing$. This arises when a multi-hop query resolves an intermediate entity at one time and reasons about it at another — finding who captained a side in one year and asking about that person's club in a different year. Per-claim verification cannot detect it, because no single claim is wrong.

4 TIMEGUARD

The mechanisms in Section 3.4 share a shape: the agent has the information it needs, or could get it cheaply, but carries no structured representation of what has been temporally established and what has not. If stale grounding stems from missing state rather than missing evidence, supplying the state during execution should change behaviour.

TIMEGUARD is an inference-time tracker running alongside the agent. After step t completes with (ρ_t, a_t, o_t), TIMEGUARD extracts temporal expressions from o_t, updates the intervals V for the claims o_t bears on, recomputes A from Equation (2) and W from Equation (3), and emits a compact *temporal state block* $\mathcal{T}_{t+1}$: the anchor $\tau^\star$; one line per claim giving its interval and alignment; the running intersection W; and the list of claims that remain UNSCOPED. The agent continues from the augmented context,

$$\rho_{t+1}, a_{t+1} \sim \pi\big(\cdot \mid \mathcal{H}_t, \mathcal{T}_{t+1}\big). \tag{5}$$

Two design choices matter. First, TIMEGUARD updates only the evidence side of the state: it writes V and A, never B. The agent's temporal stance must remain the agent's, or the intervention would manufacture the very belief we are trying to measure. Second, the block summarises what has been retrieved; it is not an instruction about what to do with it. TIMEGUARD never tells the agent to search again, reject a candidate or prefer recency — it reports the geometry and leaves the decision to π, which keeps it an orthogonal, plug-and-play module.

For prompt-based backbones π maintains $\mathcal{T}$ itself. For post-trained agents we use a separate updater $\mathcal{S}$ and concatenate its output into the observation, $o'_t = o_t \oplus \mathcal{T}_{t+1}$, because a policy trained on a fixed trajectory format destabilises when asked to emit an unfamiliar structured block, whereas treating the block as part of the environment's return leaves that format intact. Details and prompts are in Sections C and H.

5 EXPERIMENTS

5.1 SETUP

Evaluation set. We assemble TAQS, a set of 268 temporally anchored queries drawn from existing benchmarks that already involve long-horizon information seeking or time-sensitive knowledge: BrowseComp (Wei et al., 2025), FRAMES (Krishna et al., 2025), FreshQA (Vu et al., 2024), SituatedQA (Zhang & Choi, 2021), TimeQA (Chen et al., 2021) and WebWalkerQA (Wu et al., 2025b). An instance is admitted only if (i) its anchor resolves to at least month granularity, (ii) the gold answer changes under a counterfactual shift of the anchor, and (iii) the question decomposes into at least two claims that must hold jointly at the anchor. Criterion (ii) is what makes the set a stress test rather than a sample of ordinary questions: where the answer is time-invariant, stale grounding is unmeasurable. Composition, anchor types and the selection prompts are in Sections A and H.

Baselines. We evaluate RL-trained search agents: **Search-R1** (Jin et al., 2025) and **R1-Searcher** (Song et al., 2025), which incentivise multi-turn search through outcome rewards; **Hybrid Deep Searcher** (HDS) (Ko et al., 2026), which mixes parallel and sequential search; **WebSailor** (Li et al., 2025b) and **WebExplorer** (Liu et al., 2025), trained for long-horizon navigation; **DR Tulu** (Shao et al., 2026), trained with evolving rubrics; and **Tongyi DeepResearch** (TongyiDR) (Li et al., 2025a). We also evaluate the prompt-based scaffolds **Search-o1** (Li et al., 2025c) and **ReAct** (Yao et al., 2023) on `gpt-oss-20b` and `gpt-oss-120b` (Agarwal et al., 2025). To separate temporal state from simply thinking longer we add a test-time scaling control, **ReAct + TTS**, adapting the budget-forcing procedure of Muennighoff et al. (2025) to search agents by suppressing the answer token and appending a continuation cue.

Metrics. **Accuracy (Acc, ↑)** is the proportion of instances answered correctly, judged by a strong model (Singh et al., 2026) following standard practice for open-ended agentic benchmarks (Zheng et al., 2023). **Stale-grounding rate (SGR, ↓)** is the proportion of answers satisfying Equation (4). **Temporal consistency score (TCS, ↑)** is the mean over instances of the fraction of supporting claims that are ALIGNED; being graded, it distinguishes an answer with one unanchored claim from one with none anchored. **Anchored interval coverage (AIC, ↑)** is the mean fraction of supporting claims for which the trajectory establishes any determined interval, and so measures whether the agent gathers temporally scoped evidence at all, independently of whether that scope covers the anchor.

Evaluation details. CHRONOSCOPE is maintained by an evaluator model $\mathcal{M}$ (`gpt-oss-120b`); we validate its annotations against human judgements on a stratified sample and observe substantial to almost perfect agreement on all four subtasks (Section F). All agents share the same tools, a cap of 100 turns and identical retrieval settings (Section B).

5.2 RESULTS

Stale grounding is pervasive, and accuracy conceals it. Table 1 reports Acc, SGR and TCS across all systems. Accuracy varies by a factor of five between the weakest and strongest baseline, and SGR falls as accuracy rises — stronger agents do anchor more of their evidence. But the relationship is far from sufficient: TongyiDR, the strongest baseline, still leaves 55.6% of its answers stale-grounded, and its TCS of 53.1 means roughly half of all supporting claims are never tied to the anchor. Search-R1 and R1-Searcher answer about one instance in ten correctly while grounding nine in ten answers on evidence that does not cover the anchor — what one would expect of a system that treats retrieval as topical matching rather than temporal verification.

Stale grounding persists inside correct answers. Figure 3 stratifies each system by correctness and grounding status. Incorrect answers are overwhelmingly stale-grounded, as expected; the informative

Method	Backbone	Acc (↑)	SGR (↓)	TCS (↑)
Reinforcement-learning-trained search agents				
Search-R1 (Jin et al., 2025)	Qwen2.5-7B	11.2	89.9	21.8
R1-Searcher (Song et al., 2025)	Qwen2.5-7B	9.7	91.0	19.5
Hybrid Deep Searcher (Ko et al., 2026)	Qwen2.5-7B	18.7	83.2	29.3
DR Tulu (Shao et al., 2026)	Llama-3.1-8B	21.6	80.6	31.5
WebSailor (Li et al., 2025b)	Qwen3-8B	26.1	78.7	33.6
WebExplorer (Liu et al., 2025)	Qwen3-8B	34.7	71.3	39.1
TongyiDR (Li et al., 2025a)	30B-A3B	51.5	55.6	53.1
TongyiDR + TIMEGUARD **(Ours)**	30B-A3B	**55.2**	41.8	73.3
Prompt-based scaffolds				
Search-o1 (Li et al., 2025c)	gpt-oss-20b	20.9	82.1	26.1
ReAct (Yao et al., 2023)	gpt-oss-20b	31.7	69.4	37.4
ReAct + TIMEGUARD **(Ours)**	gpt-oss-20b	37.3	52.6	62.7
Search-o1 (Li et al., 2025c)	gpt-oss-120b	29.9	73.5	32.2
ReAct (Yao et al., 2023)	gpt-oss-120b	35.4	68.3	36.0
ReAct + TTS (Muennighoff et al., 2025)	gpt-oss-120b	38.1	59.7	46.5
ReAct + TIMEGUARD **(Ours)**	gpt-oss-120b	46.6	**40.7**	**75.7**

Table 1: Main results on TAQS. Acc is answer accuracy, SGR the stale-grounding rate of Equation (4), and TCS the graded temporal consistency score. Arrows give the preferred direction; best value per column in bold. TIMEGUARD improves all three metrics for every scaffold it is applied to, and the improvement in SGR and TCS is larger than the improvement in accuracy — the intervention changes how answers are supported, not only whether they are right.

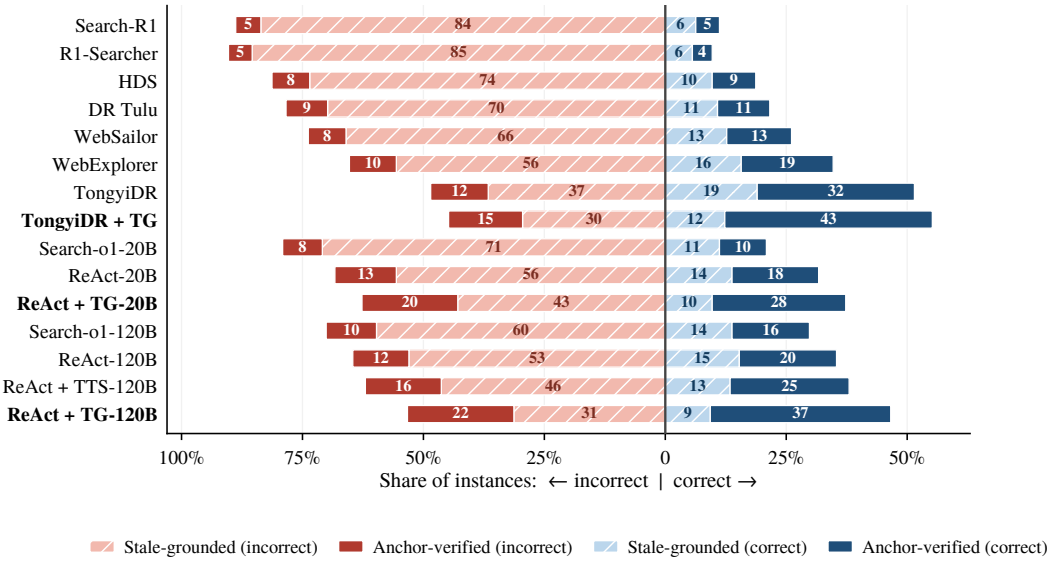


Figure 3: Breakdown of answers by correctness and grounding status. Hatched bars are stale-grounded. Systems augmented with TIMEGUARD are shown in bold. TIMEGUARD converts stale-grounded correct answers into anchor-verified ones and moves a substantial share of stale-grounded incorrect answers across the centre line.

region is right of the centre line. Even for TongyiDR, 19 of every 100 instances are answered correctly on evidence that does not cover the anchor, against 32 with anchored evidence; for ReAct-120B the split is 15 against 20, so nearly half of its correct answers are unsupported at the anchor.[1] Two systems

[1]Manual inspection of anchor-verified but incorrect cases indicates a substantial fraction are annotation artefacts rather than agent errors, principally gold answers themselves anchored to a different moment; see Section E.

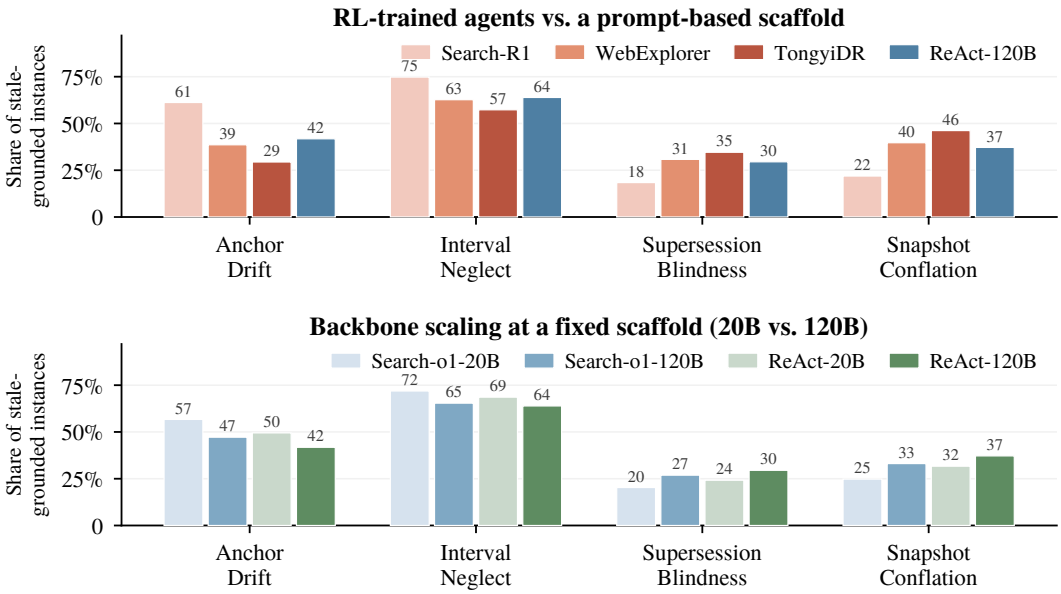


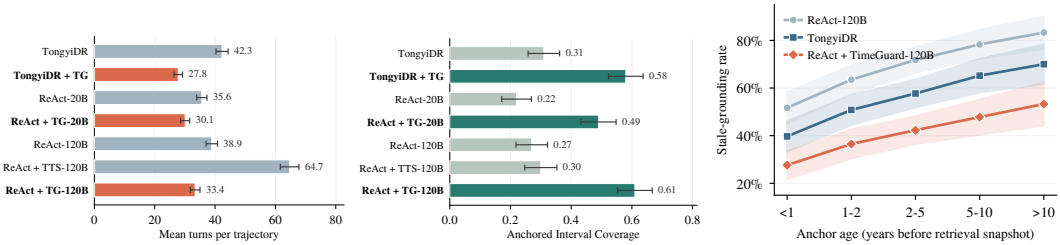
Figure 4: Distribution of the four stale-grounding mechanisms, as a share of each system's stale-grounded instances. Mechanisms may co-occur, so shares need not sum to one. The top panel contrasts RL-trained agents with a prompt-based scaffold; the bottom isolates backbone scale at a fixed scaffold. Capability reduces anchor drift and interval neglect but increases supersession blindness and snapshot conflation.

can therefore report the same accuracy while differing sharply in how much of it is load-bearing, and that difference is what shows up as silent degradation when the underlying facts change.

Training and scale redistribute the mechanisms rather than removing them. Figure 4 shows how the four mechanisms are distributed: the top panel contrasts trained agents with a prompt-based scaffold, the bottom isolates backbone scale at a fixed scaffold. Both interventions consistently reduce anchor drift and interval neglect — stronger agents restate the anchor more faithfully and are likelier to look for a date — and both consistently *increase* supersession blindness and snapshot conflation. Moving from Search-o1-20B to Search-o1-120B lowers anchor drift from 56.7% to 47.2% of stale-grounded instances while raising snapshot conflation from 24.8% to 33.1%; the same holds for ReAct and across the trained agents, where TongyiDR shows the lowest anchor drift (29.4%) and the highest snapshot conflation (46.2%) of any baseline. The explanation is mechanical: stronger agents run longer trajectories with more hops, so they retrieve more updates they can fail to adopt and compose more intermediate entities that can be resolved at inconsistent times. Capability buys temporal literacy at the level of single claims and creates new opportunities for failure at the level of chains.

TIMEGUARD improves accuracy and temporal consistency across scaffolds. Adding TIMEGUARD improves every system it is applied to (Table 1). Accuracy rises by 3.7 points for TongyiDR, 5.6 for ReAct-20B and 11.2 for ReAct-120B, while SGR falls by 13.8, 16.8 and 27.6 points. TCS moves furthest, from 53.1 to 73.3 for TongyiDR and from 36.0 to 75.7 for ReAct-120B, so the gain is a broad shift in how much of each supporting chain gets tied to the anchor rather than a handful of instances flipping. ReAct + TTS also reduces SGR, from 68.3 to 59.7, but Figure 3 shows it does so mostly by anchoring answers the system already had right, without moving many instances across the centre line. TIMEGUARD does both, and the accuracy gap between the two on the same backbone is 8.5 points.

Decomposing the improvement by mechanism (Section D) shows the gain is uneven. Anchor drift and interval neglect fall sharply everywhere — for ReAct-120B, from 41.8% to 14.2% and from 63.9% to 28.5% of stale-grounded instances — as one expects of a module that restates the anchor every turn and names the claims that remain unscoped. Snapshot conflation responds least, falling only from 37.2% to 26.3% and rising slightly on ReAct-20B. Reporting an empty joint window tells the agent its chain cannot be repaired by one more search; acting on that requires revisiting an earlier



(a) Trajectory length. **(b)** Interval coverage. **(c)** Effect of anchor age.

Figure 5: Where the gain comes from. (a) TIMEGUARD shortens trajectories while test-time scaling lengthens them. (b) TIMEGUARD roughly doubles the share of supporting claims that carry a determined validity interval. (c) Stale grounding worsens as the anchor recedes from the retrieval snapshot, and the margin TIMEGUARD provides widens with it. Bars and bands are bootstrap standard errors.

hop, which the smaller backbone largely does not do. Chain-level temporal reasoning is therefore not solved by exposing chain-level state, and we regard it as the harder residual problem.

Explicit state makes search shorter, and helps most where the web is weakest. A tracker that runs every turn could plausibly pay for accuracy with latency; it does not (Figure 5a). TongyiDR falls from 42.3 to 27.8 turns with TIMEGUARD, ReAct-120B from 38.9 to 33.4, ReAct-20B from 35.6 to 30.1, while the test-time scaling control moves the other way, from 38.9 to 64.7 turns, for a smaller accuracy gain. Without temporal state, agents re-issue queries they have already answered because they cannot tell which claims are settled; a running list of unscoped claims gives the search a stopping condition. The mechanism behind the gain is visible in Figure 5b: TIMEGUARD roughly doubles anchored interval coverage — 0.31 to 0.58 for TongyiDR, 0.22 to 0.49 for ReAct-20B, 0.27 to 0.61 for ReAct-120B — while test-time scaling barely moves it, from 0.27 to 0.30. Whatever extra reasoning the TTS control performs, it is not the work of dating evidence. Finally, Figure 5c stratifies instances by anchor age. Every system degrades as the anchor recedes from the retrieval snapshot, the signature of a present-biased retrieval stack: ReAct-120B climbs from 51.7% SGR on anchors under a year old to 83.3% on those more than a decade old. TIMEGUARD degrades along the same slope from a lower base, and the margin it provides grows from 24.1 to 30.0 points.

6 CONCLUSION

We asked whether search agents establish that their evidence holds at the time their question asks about. On temporally anchored queries they frequently do not: agents commit to answers whose supporting evidence never covers the anchor, a failure we call stale grounding and one that final-answer accuracy does not reveal. To diagnose it we introduced CHRONOSCOPE, which makes evidential support interval-valued and compares it against a query-side anchor, and used it to isolate four recurring mechanisms — anchor drift, interval neglect, supersession blindness and snapshot conflation. We then showed that exposing temporal state during execution, through TIMEGUARD, reduces stale grounding, improves accuracy and shortens trajectories across both trained and prompt-based agents.

Two limitations point directly at future work. Snapshot conflation resists the intervention, which suggests that chain-level consistency needs a mechanism for revising an earlier hop rather than only a report that the chain is inconsistent. And TIMEGUARD acts purely at inference time, while the support maps it produces are exactly the supervision a training procedure would need — which makes post-training an agent for temporal verification a natural next step.

AI STATEMENT

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, and Edwin Arbus. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2108.06314.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Bahare Fatemi, Seyed Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.09170.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971. doi: 10.1037/h0031619.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. *arXiv preprint arXiv:2508.07976*, 2025. URL https://arxiv.org/abs/2508.07976.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2web 2: Evaluating agentic search with agent-as-a-judge. In *Advances in Neural Information Processing Systems 39, Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2506.21506.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. Complextempqa: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9100–9112, 2025. doi: 10.18653/v1/2025.emnlp-main.463. URL https://aclanthology.org/2025.emnlp-main.463/.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. Temptabqa: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2431–2453, 2023. doi: 10.18653/v1/2023.emnlp-main.149. URL https://aclanthology.org/2023.emnlp-main.149/.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. Characterizing deep research: A benchmark and formal definition. *arXiv preprint arXiv:2508.04183*, 2025. URL https://arxiv.org/abs/2508.04183.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pp. 1057–1062, 2018. doi: 10.1145/3184558.3191536.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. Tiq: A benchmark for temporal question answering with implicit time constraints. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1394–1399, 2024. doi: 10.1145/3589335.3651895.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2503.09516.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2207.13332.

Wonjoong Kim, Sangwu Park, Yeonjun In, Sein Kim, Dongha Lee, and Chanyoung Park. Beyond the final answer: Evaluating the reasoning trajectories of tool-augmented agents. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.02837.

Dayoon Ko, Jihyuk Kim, Haeju Park, Sohyeon Kim, Dahyun Lee, Yongrae Jo, Gunhee Kim, Moontae Lee, and Kyungjae Lee. Hybrid deep searcher: Scalable parallel and sequential search reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.19113.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, 1977. doi: 10.2307/2529310.

Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, and Yong Jiang. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025a. URL https://arxiv.org/abs/2510.24701.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025b. URL https://arxiv.org/abs/2507.02592.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, 2025c. doi: 10.18653/v1/2025.emnlp-main.276. URL https://aclanthology.org/2025.emnlp-main.276/.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. In *Advances in Neural Information Processing Systems 38*, 2025d. URL https://arxiv.org/abs/2504.21776.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark

for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Junteng Liu, Yunji Li, Chi Zhang, Jingyang Li, Aili Chen, Ke Ji, Weiyu Cheng, Zijia Wu, Chengyu Du, Qidi Xu, Jiayuan Song, Zhengmao Zhu, Wenhu Chen, Pengyu Zhao, and Junxian He. Webexplorer: Explore and evolve for training long-horizon web agents. *arXiv preprint arXiv:2509.06501*, 2025. URL https://arxiv.org/abs/2509.06501.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5944–5958, 2022. doi: 10.18653/v1/2022.naacl-main.435. URL https://aclanthology.org/2022.naacl-main.435/.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, 2025. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6036–6063, 2025. doi: 10.18653/v1/2025.acl-long.301. URL https://aclanthology.org/2025.acl-long.301/.

Bhawna Piryani, Abdelrahman Abdallah, Jamshid Mozafari, Avishek Anand, and Adam Jatowt. It's high time: A survey of temporal information retrieval and question answering. *arXiv preprint arXiv:2505.20243*, 2025. URL https://arxiv.org/abs/2505.20243.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*, 2024. URL https://arxiv.org/abs/2408.08067.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2511.19399.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, and Ahmed El-Kishky. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2026. URL https://arxiv.org/abs/2601.03267.

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025. URL https://arxiv.org/abs/2501.09136.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025. URL https://arxiv.org/abs/2503.05592.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3025–3035, 2022. doi: 10.1145/3477495.3531734.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.00935.

Yuqing Wang and Yun Zhao. Tram: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6389–6415, 2024. doi: 10.18653/v1/2024.findings-acl.382. URL https://aclanthology.org/2024.findings-acl.382/.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025. URL https://arxiv.org/abs/2504.12516.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1434–1447, 2023. doi: 10.18653/v1/2023.findings-emnlp.100. URL https://aclanthology.org/2023.findings-emnlp.100/.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhenglin Wang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Xiangru Tang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. Webdancer: Towards autonomous information seeking agency. In *Advances in Neural Information Processing Systems 38*, 2025a. URL https://arxiv.org/abs/2505.22648.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. Webwalker: Benchmarking llms in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, 2025b. doi: 10.18653/v1/2025.acl-long.508. URL https://aclanthology.org/2025.acl-long.508/.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13300.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Zhichao Xu, Zongyu Wu, Yun Zhou, Aosong Feng, Kang Zhou, Sangmin Woo, Kiran Ramnath, Yijun Tian, Xuan Qi, Weikang Qiu, Lin Lee Cheong, and Haibo Ding. Beyond correctness: Rewarding faithful reasoning in retrieval-augmented generation. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2510.13272.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024. URL https://arxiv.org/abs/2401.15884.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8938–8968, 2024. doi: 10.18653/v1/2024.emnlp-main.505. URL https://aclanthology.org/2024.emnlp-main.505/.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14672–14685, 2024. doi: 10.18653/v1/2024.emnlp-main.813. URL https://aclanthology.org/2024.emnlp-main.813/.

Michael J.Q. Zhang and Eunsol Choi. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, 2021. doi: 10.18653/v1/2021.emnlp-main.586. URL https://aclanthology.org/2021.emnlp-main.586/.

Michael J.Q. Zhang and Eunsol Choi. Mitigating temporal misalignment by discarding outdated facts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14213–14226, 2023. doi: 10.18653/v1/2023.emnlp-main.879. URL https://aclanthology.org/2023.emnlp-main.879/.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A. Smith. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15015–15040, 2024. doi: 10.18653/v1/2024.findings-acl.892. URL https://aclanthology.org/2024.findings-acl.892/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, and Yonghao Zhuang. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025. doi: 10.18653/v1/2025.emnlp-main.22. URL https://aclanthology.org/2025.emnlp-main.22/.

A Construction of TAQS

A.1 Admission criteria

TAQS is assembled from instances of existing benchmarks rather than written from scratch, so the queries retain the retrieval difficulty of the source tasks. Each candidate passes three filters.

(i) Anchor resolvability. The anchor must be recoverable at month granularity or finer. Explicit dates and periods resolve directly; deictic anchors resolve to the retrieval snapshot; event-relative anchors resolve to the date of the referenced event, which the question must itself determine uniquely. Questions with genuinely ambiguous anchors ("recently") are discarded.

(ii) Temporal sensitivity. The gold answer must change under a counterfactual shift of the anchor. We ask an extraction model for the answer that would be correct at two shifted anchors, one earlier and one later, and keep the instance only if at least one differs from the gold answer. Without this filter the set would not be a stress test: where the answer is time-invariant, every grounding is trivially anchor-valid.

(iii) Compositional support. The question must decompose into at least two atomic claims holding jointly at the anchor. Single-claim questions admit anchor drift and interval neglect but cannot exhibit snapshot conflation, and we want all four mechanisms measurable on the same set.

A.2 Composition

Source benchmark	n	**Share**
BrowseComp (Wei et al., 2025)	84	31.3%
FRAMES (Krishna et al., 2025)	57	21.3%
FreshQA (Vu et al., 2024)	46	17.2%
SituatedQA (Zhang & Choi, 2021)	38	14.2%
TimeQA (Chen et al., 2021)	25	9.3%
WebWalkerQA (Wu et al., 2025b)	18	6.7%
Total	268	100.0%

Table 2: Composition of TAQS.

Anchor type	n	**Claims**	n
Explicit date	97	2 claims	61
Explicit period	74	3 claims	88
Deictic present	61	4 claims	74
Event-relative	36	5 or more	45
Total	268	Total	268

Table 3: Distribution of anchor types (left) and of the number of atomic claims per instance (right).

Table 2 reports the composition of the set. BrowseComp contributes the largest share, reflecting the density of date-qualified constraints in its questions; FreshQA and SituatedQA contribute deictic anchors, TimeQA long-range explicit ones. All questions are in English. Table 3 gives the distribution of anchor types and claim counts.

Four abbreviated cases illustrate the filters (entity names paraphrased). *Admitted*: "Which club did the captain of the 2016 national side play for as of March 2021?" — anchor 2021-03, answer differs at 2018-03, three claims. *Rejected by (i)*: "Which club did that captain recently join?" — no anchor at month granularity. *Rejected by (ii)*: "In which city was the 2016 side's captain born?" — anchor resolvable, answer time-invariant. *Rejected by (iii)*: "Who led the organisation in 2019?" — a single claim.

B EXPERIMENTAL DETAILS

Trained baselines are evaluated at the checkpoints released by their authors with the hyperparameters those authors specify; prompt-based scaffolds and the CHRONOSCOPE evaluator use the open-weight `gpt-oss` family (Agarwal et al., 2025). The TIMEGUARD updater matches the size class of the agent it augments, except for post-trained agents, where it uses the 20b model for throughput. All agents share one search tool returning the top ten snippets per query and one browse tool returning at most 8,000 characters of page text, and the retrieval snapshot is held fixed across systems so anchor age is comparable: no system can retrieve a document published after the snapshot. Serving uses a paged-attention engine (Kwon et al., 2023) with asynchronous batching. The full evaluation, including the CHRONOSCOPE annotation pass over every trajectory, corresponds to roughly 1,950 accelerator-hours on 40 GB-class devices, dominated by the 120b configurations. Table 4 lists the inference settings.

Setting	**Policy** π	$\mathcal{M}$ / $\mathcal{S}$
Temperature	1.0	0.2
Top-p	1.0	1.0
Max new tokens per turn	8,192	1,024
Max turns per trajectory	100	—
Reasoning effort	high	high
Search results per call	10 snippets	—
Browse budget per call	8,000 ch.	—
Max concurrent sequences	16	16
Time granularity	—	month
Interval merge rule	—	most specific

Table 4: Inference settings. The evaluator is decoded near-greedily so repeated annotation passes over one trajectory are stable.

Component	**k-tok / inst.**	**k-tok / turn**	**Over-head**
ReAct-120B (agent)	46.3	1.19	—
ReAct + TIMEGUARD-120B (agent)	39.8	1.19	—
TIMEGUARD updater (20b)	7.8	0.23	+19.6%
TIMEGUARD updater (120b)	9.1	0.27	+22.9%
CHRONOSCOPE offline pass (120b)	11.4	0.34	n/a

Table 5: Token cost. The updater adds about a fifth to the agent's own consumption, but removes turns: end-to-end use rises from 46.3 to 48.9 k-tokens per instance, under six per cent over the baseline.

C IMPLEMENTATION OF TIMEGUARD

C.1 THE TEMPORAL STATE BLOCK

TIMEGUARD emits a fixed-format block after each agent step; the fixed format lets the agent learn the layout within one trajectory and keeps the block from crowding out retrieved text. For the running example of Figure 2:

```
[TEMPORAL STATE]
anchor:  2021-03
c1 captain of the 2016 side    [2015-08, 2017-05]    STALE
c2 club membership             [2014-11, 2019-07]    STALE
c3 club's division             [2019-01,        )    ALIGNED
joint window:  EMPTY (c1 x c2 disjoint from c3)
unscoped claims:  none
```

Three fields carry most of the signal. Per-claim alignment tells the agent which specific claim is not anchored, rather than that something is wrong; the joint window turns a set of individually plausible facts into one verdict about the chain; and the unscoped list names claims for which no date has been found at all, the cue that matters most for interval neglect.

C.2 INTEGRATION

For a prompt-based scaffold the agent maintains the block itself, so the history interleaves reasoning, actions, observations and blocks, with $\mathcal{T}_{t+1} \sim \pi(\cdot \mid \rho_t, a_t, o_t)$. For a post-trained agent a separate updater produces the block and it is folded into the observation,

$$\mathcal{T}_{t+1} \sim \mathcal{S}(\cdot \mid \rho_t, a_t, o_t), \qquad o'_t = o_t \oplus \mathcal{T}_{t+1}, \tag{6}$$

preserving the trajectory format the policy was trained on. The distinction is necessary rather than cosmetic: asking a post-trained policy to emit the block inline degraded its search behaviour well before it affected its answers, producing truncated queries and premature termination.

C.3 Cost

Table 5 reports the additional token cost. Online TimeGuard is cheaper per instance than the offline Chronoscope pass, because the updater sees one turn at a time and need not reconstruct alternative intervals for supersession detection; it adds roughly a fifth to the agent's own consumption while removing turns, so wall-clock cost is close to neutral.

D Additional Experiments

D.1 Test-time scaling as a control

Method (120b)	Acc (↑)	SGR (↓)	TCS (↑)	Turns (↓)
ReAct	35.4	68.3	36.0	38.9
ReAct + TTS	38.1	59.7	46.5	64.7
ReAct + TimeGuard **(Ours)**	**46.6**	**40.7**	**75.7**	**33.4**

Table 6: Test-time scaling (Muennighoff et al., 2025) as a control on the gpt-oss-120b ReAct scaffold (Yao et al., 2023). Extending the reasoning budget reduces stale grounding, but costs turns rather than saving them.

Variant (ReAct-120B)	Acc	SGR	TCS
No temporal state	35.4	68.3	36.0
+ anchor restatement	38.4	62.5	44.2
+ interval extraction	41.0	52.8	56.0
+ joint-window intersection	44.4	45.5	66.9
+ unscoped flag (full TimeGuard)	**46.6**	**40.7**	**75.7**

Table 7: Component ablation of the temporal state block; each row adds one field to the block emitted after every turn.

Table 6 isolates the comparison between TimeGuard and test-time scaling on one backbone; the corresponding grounding split for these three systems is the last block of Figure 3. Test-time scaling raises the anchor-verified correct share from 20% to 25% of instances while the correct total moves only from 35% to 38%: almost all of its benefit is re-verification of answers the system already had. TimeGuard raises the anchor-verified correct share to 37% and the correct total to 47%, and is the only one of the three that meaningfully shrinks the stale-grounded incorrect region, from 53% to 31%.

D.2 Effect of TimeGuard on each mechanism

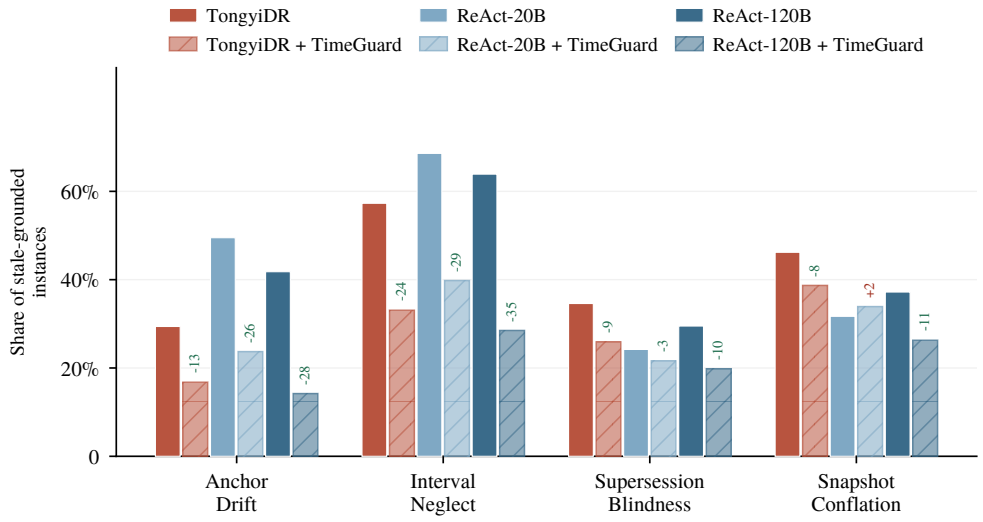


Figure 6: Effect of TimeGuard on each mechanism. Solid bars are the baseline, hatched bars the same system with TimeGuard; annotations give the signed change in percentage points. Gains concentrate on anchor drift and interval neglect; snapshot conflation is the residual.

Figure 6 gives the per-mechanism decomposition summarised in Section 5.2. The two mechanisms the block addresses directly — anchor drift, through the restated anchor, and interval neglect, through the unscoped-claim list — fall by wide margins on all three systems, and widest on the strongest backbone. Supersession blindness improves modestly and uniformly: naming the interval that supports a claim makes a later, better-dated interval easier to notice, but nothing instructs the agent to prefer it. Snapshot conflation is the exception, falling on TongyiDR and ReAct-120B but rising on ReAct-20B, from 31.7% to 33.9%; the transition analysis below shows why.

D.3 Which part of TimeGuard matters

Table 7 adds the components of the temporal state block one at a time to ReAct-120B. Restating the anchor every turn is the cheapest component and already removes a fifth of the SGR gap, consistent with anchor drift being partly an attention failure rather than a knowledge failure. Interval extraction is the largest single contributor, adding 9.7 points of SGR reduction on top of anchor restatement and 11.8 points of TCS. The joint intersection helps mainly on multi-hop instances, and the unscoped-claim flag contributes the remaining 4.8 points — a small component with an outsized effect on interval neglect, because it turns a silent omission into an explicit item on the agent's list.

D.4 Per-benchmark breakdown

		Source benchmark					
Method	**Metric**	**BrowseComp**	**FRAMES**	**FreshQA**	**SituatedQA**	**TimeQA**	**WebWalkerQA**
ReAct-120B	Acc (↑)	22.5	38.6	39.1	44.7	48.0	38.9
	SGR (↓)	73.8	66.7	69.6	63.2	60.0	66.7
TongyiDR	Acc (↑)	36.9	54.4	56.5	63.2	64.0	55.6
	SGR (↓)	60.7	54.4	56.5	50.0	48.0	55.6
ReAct + TimeGuard-120B	Acc (↑)	33.2	49.1	50.0	57.9	60.0	50.0
	SGR (↓)	46.5	40.4	41.3	34.2	32.0	38.9

Table 8: Per-benchmark accuracy and stale-grounding rate. Weighted by the subset sizes of Table 2, each row reproduces the corresponding aggregate in Table 1.

Table 8 splits accuracy and SGR by source benchmark. Stale grounding is highest on BrowseComp, whose questions chain the most constraints and whose anchors are most often deeply historical, and lowest on TimeQA, whose anchors are usually stated outright. The ordering of the three systems holds on every subset, so the effect is not an artefact of one source.

D.5 How failure modes move

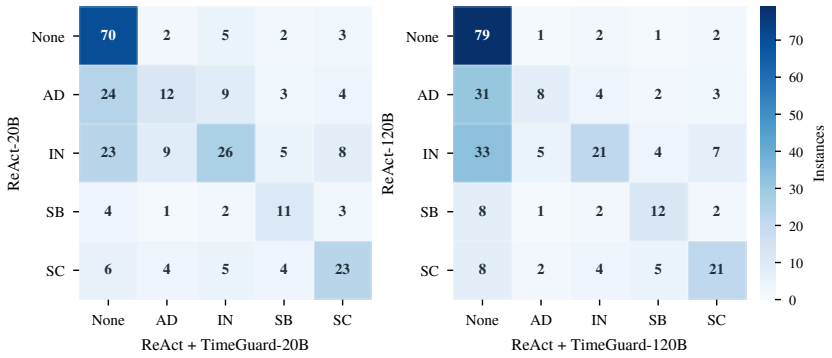


Figure 7: Transition of the dominant failure mechanism for individual instances before and after adding TimeGuard (AD: anchor drift; IN: interval neglect; SB: supersession blindness; SC: snapshot conflation; None: no mechanism fires). Rows are the baseline mechanism, columns the mechanism after integration.

Figure 7 tracks individual instances rather than aggregate rates. The dominant flow in both panels is into the *None* column: 70 instances on the 20b backbone and 79 on the 120b backbone show no

mechanism either before or after, and a further 24 and 31 move out of anchor drift into *None*. The instructive cells are off-diagonal: on the 20b backbone, 9 instances move from anchor drift into interval neglect and 4 into snapshot conflation, so told the anchor is elsewhere, the smaller model stops resolving at the present but does not reliably go and date the claim. The 120b backbone shows less of this exchange into interval neglect or snapshot conflation — 7 of its 48 anchor-drift instances against 13 of 52 — mirroring the aggregate result that TIMEGUARD composes better with a stronger policy.

E MANUAL REVIEW OF ANCHOR-VERIFIED BUT INCORRECT ANSWERS

Category	Cases	Share
Gold answer anchored elsewhere	26	44.1%
Reasoning error unrelated to time	15	25.4%
Ambiguous or under-specified anchor	8	13.6%
Interval from an unreliable source	7	11.9%
Annotation error by the evaluator	3	5.1%
Total reviewed	59	100.0%

Table 9: Manual review of the anchor-verified but incorrect answers of ReAct + TIMEGUARD-120B. Fewer than half are agent errors.

Annotation subtask	Mean agr.	Maj. agr.	κ
Anchor identification	0.95	0.98	0.81
Validity-interval extraction	0.89	0.95	0.74
Alignment labelling	0.92	0.97	0.78
Mechanism attribution	0.86	0.93	0.69

Table 10: Agreement between CHRONOSCOPE and three human annotators on 60 stratified instances; κ is Fleiss' κ among annotators.

The cells of Figure 3 hardest to interpret are the anchor-verified incorrect ones: the evidence covers the anchor and the answer is still judged wrong. We reviewed all 59 such cases for ReAct + TIMEGUARD-120B, inspecting the question, the gold and predicted answers, and the intervals CHRONOSCOPE recorded, and asking whether the recorded evidence entails the prediction at the anchor.

Table 9 gives the outcome. The largest category is gold answers themselves anchored to a different moment than the question states, usually because the source benchmark was annotated when a different answer was current — the annotation is stale in exactly the way we are measuring. Only three cases were evaluator errors, where CHRONOSCOPE recorded an interval the retrieved text did not support; the low rate is consistent with the agreement figures in Section F. The anchor-verified incorrect region is therefore a mixture in which agent error is the minority, so the true accuracy ceiling on TAQS sits somewhat above what the gold labels admit.

F HUMAN EVALUATION OF CHRONOSCOPE ANNOTATIONS

Because CHRONOSCOPE is model-maintained, its annotations must themselves be validated. We sampled 60 instances stratified by system and by whether the answer was flagged stale-grounded, and had three annotators independently label four subtasks: which anchor the question carries, what interval a given piece of retrieved evidence supports, whether that interval aligns with the anchor, and which mechanism the trajectory exhibits. Annotators saw the question, the retrieved documents and the prediction, but not the evaluator's output.

Table 10 reports agreement with the evaluator and Fleiss' κ (Fleiss, 1971) among annotators. Anchor identification is the most reliable subtask, as expected for something largely stated in the question. Mechanism attribution is the least reliable, at $\kappa = 0.69$ — substantial but not almost-perfect by the conventional reading (Landis & Koch, 1977) — with disagreement concentrated on trajectories where two mechanisms co-occur and annotators differ on which dominates. This limits the mechanism-level analysis specifically, not the detection criterion of Equation (4), whose components agree at $\kappa = 0.74$ and 0.78.

G QUALITATIVE EXAMPLES

Example 1 — a stale-grounded trajectory (ReAct-120B)

Question. As of June 2022, which institution employed the researcher who had directed the Meridian Observatory during its first survey?

Gold. Farrow Institute of Technology **Prediction.** Ashgrove University ✗

ASSISTANT *think* First find who directed the Meridian Observatory during its first survey.
search `Meridian Observatory first survey director`

TOOL [1] "The first survey ran from 2009 to 2013 under the direction of L. Varga . . . " [2] "Varga, then at Ashgrove University, coordinated the instrument team . . . "

ASSISTANT *think* So Varga directed it, and Varga is at Ashgrove. Let me confirm the affiliation.
search `L. Varga Ashgrove University`

TOOL [1] "L. Varga (Ashgrove University) — publications, group page . . . " [2] "Varga joined the Farrow Institute of Technology in September 2020 to lead the new astrophysics centre . . . "

ASSISTANT *think* The group page confirms Ashgrove. That is consistent with the survey record.
answer Ashgrove University

CHRONOSCOPE support map. $\tau^\star = 2022\text{-}06$.

Claim	Evidence	$V(k,c)$	$A(k,c)$	$B(k,c)$
c_1 directed the first survey	search 1, doc 1	[2009-01, 2013-12]	ALIGNED†	ANCHORED
c_2 employed by the institution	search 1, doc 2	[2007-09, 2020-08]	STALE	PRESENTIST
c_3 institution is the employer at $\tau^\star$	search 2, doc 1	$\perp$	UNSCOPED	AGNOSTIC
c_2' *superseding evidence*	search 2, doc 2	[2020-09, ∞)	—	*not adopted*

† c_1 concerns a past event, so its interval is evaluated against the event rather than $\tau^\star$ (Section A).
Verdict. Stale-grounded. **Mechanisms.** supersession blindness (c_2' retrieved and ignored) and interval neglect (c_3 never dated): an undated group page outranked a dated transfer announcement.

Example 2 — the same query with TIMEGUARD (ReAct + TIMEGUARD-120B), anchor-verified

Question. As of June 2022, which institution employed the researcher who had directed the Meridian Observatory during its first survey?

Gold. Farrow Institute of Technology **Prediction.** Farrow Institute of Technology ✓

ASSISTANT *think* Anchor is 2022-06. Two things must hold then: who directed the first survey, and where that person worked.
search `Meridian Observatory first survey director`

TOOL [1] "The first survey ran from 2009 to 2013 under the direction of L. Varga . . . " [2] "Varga, then at Ashgrove University, coordinated the instrument team . . . "

STATE
```
[TEMPORAL STATE] anchor:  2022-06
c1 directed first survey   [2009-01, 2013-12]   ALIGNED (event-scoped)
c2 employer at anchor      [2007-09, 2013-12]   STALE
unscoped claims:  none      joint window:  EMPTY
```

ASSISTANT *think* The employer claim is stale — the evidence stops in 2013 and I need 2022. Search for the affiliation at that date, not the one attached to the survey.
search `L. Varga affiliation 2022`

TOOL [1] "Varga joined the Farrow Institute of Technology in September 2020 . . . " [2] "Directory, 2022: L. Varga, Farrow Institute of Technology . . . "

STATE
```
c2 employer at anchor    [2020-09,        )   ALIGNED     unscoped:  none
joint window:  [2020-09, ) ∋ anchor
```

ASSISTANT *think* Both claims are aligned and the joint window contains the anchor.
answer Farrow Institute of Technology

Verdict. Anchor-verified. The block did not say which document to prefer; it reported that the interval supporting c_2 ended before the anchor, and the agent chose to re-search — the typical shape of the intervention's effect in Figure 6.

H PROMPTS

Anchor extraction and temporal-sensitivity filter. This prompt implements criteria (i) and (ii) of Section A: it resolves the anchor and probes whether the answer would change if the anchor moved.

Prompt: anchor resolution and counterfactual sensitivity

```
### Task
You will be given a natural-language question. Determine the point in time the question
asks about, and decide whether the correct answer depends on it.
### Definitions
Anchor -- the instant or interval at which the question's conditions must hold:
EXPLICIT_DATE ("as of 3 August 2024"); EXPLICIT_PERIOD ("during the 2019 season"); DEICTIC
("the current chair") - resolve to {snapshot_date}; EVENT_RELATIVE ("at the time the
treaty was signed") - resolve to the date of the referenced event, which the question must
uniquely determine.
Not an anchor -- vague recency ("recently", "lately"), or a date that only names an entity
("the 2016 Olympics logo").
### Instructions
1. Resolve the anchor to month granularity or finer; if you cannot, output "anchor":
null and stop.
2. Give the answer that would be correct at the anchor shifted three years earlier and
three years later. Write "unknown" if you cannot determine one.
3. Set "time_sensitive" to true if either shifted answer differs from the answer at the
anchor.
4. Do not attempt to solve the question beyond what step 2 requires.
### Output (JSON only)
{"anchor": "YYYY-MM", "anchor_type": "...", "evidence_span": "...",
 "answer_at_anchor": "...", "answer_earlier": "...", "answer_later": "...",
 "time_sensitive": true|false}
### Input
Question: {question}  Snapshot date: {snapshot_date}
```

Claim decomposition. This prompt produces the set $\mathcal{C}$ of atomic claims a candidate answer must satisfy at the anchor. Claims must be externally checkable and atomic, since a compound claim hides exactly the interval mismatch we want to detect.

Prompt: atomic claim decomposition

```
### Task
Decompose a question into the atomic claims that a correct answer must satisfy at the
given anchor.
### Rules
1. Each claim is one externally verifiable proposition about the answer entity; if it
splits into conditions that could hold at different times, split it.
2. Mark each claim ANCHOR_SCOPED (must hold at the anchor; default) or EVENT_SCOPED (a
completed past event, evaluated against that event).
3. Exclude presentation constraints and implicit world assumptions.
4. Preserve lookup chains: if resolving one entity is needed to state another claim,
emit both and record the dependency.
5. Minimum two claims; otherwise output an empty list.
### Output (JSON only)
{"claims": [{"id": "c1", "text": "...", "scope": "ANCHOR_SCOPED", "depends_on": []},
...]}
### Input
Question: {question}  Anchor: {anchor}
```

CHRONOSCOPE annotation. The core of the offline evaluation, applied turn by turn: the annotator sees the state after turn $t-1$, the tuple (ρ_t, a_t, o_t) and the next reasoning trace ρ_{t+1}, and returns the updated support map. Separating V (what the evidence licenses) from B (what the agent believes) is what makes the mechanism attribution of Section 3.4 possible, so the prompt insists on deriving them from different inputs.

Prompt: validity-interval extraction, alignment, and stance

```
### Task
Update the temporal support map for one step of a search trajectory.
### You are given
anchor, claim list, the support map after the previous step, the current tuple (reasoning,
action, observation), and the agent's next reasoning trace.
### Step 1 -- candidates
List every answer candidate in the observation; set status ACTIVE if the next reasoning
trace selects it, REJECTED if the trace rules it out, STORED otherwise.
### Step 2 -- validity intervals (from the OBSERVATION ONLY)
For each (candidate, claim) pair the observation bears on, extract the interval over which
the retrieved text licenses the claim. Use explicit dates when present ("from 1925 to
1968", "since 2019"). A publication date alone is NOT an interval; it bounds nothing. If
the text asserts the claim with no temporal scope, output null. If this observation
```

```
supports a DIFFERENT interval than the map records, keep both and mark the newer one
supersedes.  Use only the retrieved text, never your own knowledge.
### Step 3 -- alignment (mechanical)
ALIGNED if the interval contains the anchor; STALE if it ends before; PREMATURE if it
begins after; UNSCOPED if null.
### Step 4 -- temporal stance (from the REASONING TRACE ONLY)
ANCHORED if the trace relates the claim to the anchor; PRESENTIST if it asserts the claim
as holding now or at retrieval time without relating it to the anchor; AGNOSTIC if the
trace says nothing about time for that claim.
### Output (JSON only)
{"candidates":  [{"name":  "...", "status":  "...",
 "claims":  [{"id":  "c1", "interval":  ["YYYY-MM", "YYYY-MM"|null], "supersedes":  false,
  "alignment":  "...", "stance":  "...", "quote":  "..."}]}]}
```

TIMEGUARD online updater. The online updater reuses the annotation prompt above with three modifications, so that any behavioural change is attributable to the state rather than to instruction-following.

Prompt: TIMEGUARD temporal state block (diff from the annotator)

```
### Removed
- Step 1 (candidate enumeration) and Step 4 (temporal stance).  The updater never writes
what the agent believes.
### Added
- After Step 3, recompute the joint window as the intersection of all non-null intervals;
report EMPTY if it is empty and name the disjoint pair.
- List every claim whose interval is still null.
### Hard constraints on the output
- Emit ONLY the block below:  no commentary, no suggestions, no next action.
- Never tell the agent to search again, prefer a source, or reject a candidate.  Report the
state; the agent decides.  Keep it under 12 lines.
### Format
[TEMPORAL STATE]
anchor:  {anchor}
<claim id> <short label>   [<start>, <end>]   <ALIGNED|STALE|PREMATURE|UNSCOPED>
...
joint window:  <interval or EMPTY (<ci> x <cj> disjoint)>
unscoped claims:  <comma-separated ids, or none>
```