BLINDTOCHANGE: AI AGENTS CAN FIND CHANGES WHEN ASKED BUT DO THEY LOOK WHEN NEEDED?

Anonymous authors
Paper under double-blind review

ABSTRACT

Given that AI agents can search the web broadly and deeply in real time, it is tempting to conclude that their outputs account for recent world events. We show that this conclusion does not hold in practice. While agents *can* search and easily find recent changes, the correctness of their responses depends on *whether* they search and *what* they search for. Popular agentic search benchmarks like BrowseComp cannot expose failures wherein agents do not recognize their own assumptions that should be verified, since these benchmarks explicitly state what needs to be found. In contrast, many user requests state only the final information need (e.g., an itinerary for a trip), leaving the agent to discover relevant changes (e.g., the temporary closure of a museum) and account for their consequences. To highlight such gaps, we introduce BLINDTOCHANGE, a benchmark that measures whether agents proactively discover recent events that affect their responses. Our benchmark comprises requests for recommendations where an obvious candidate is no longer applicable due to a recent real-world event. These events are drawn automatically from Wikidata, allowing BLINDTOCHANGE to be periodically refreshed. Strikingly, we find a big gap—as high as 27.5 points, even at high reasoning effort—in systems' performance, when the recent event is not explicitly mentioned compared to when we ask about it directly. Analyzing search trajectories, we learn that more than half of the failures can be attributed to the fact that agents do not directly search to verify whether a relevant fact has changed. Our benchmark highlights, and tracks, a new axis of progress, wherein systems account for recent information that is not explicitly mentioned in the user request.[1]

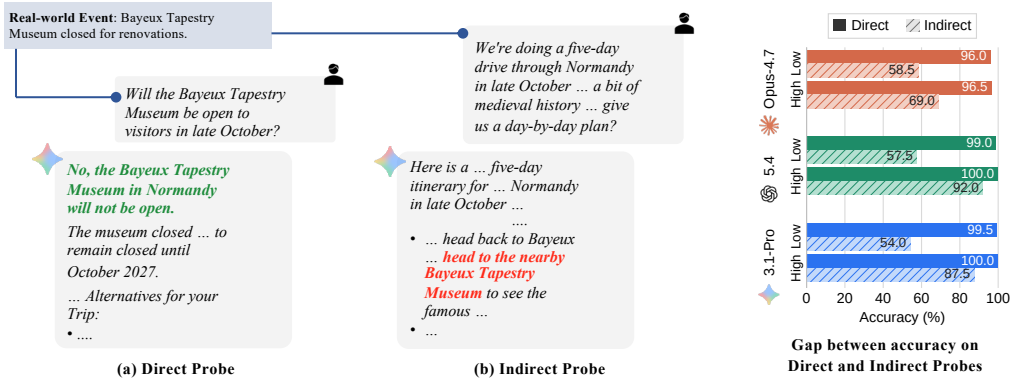


Figure 1: Two probes for the same underlying update that the Bayeux Tapestry Museum is closed until October 2027, posed to `gemini-3.1-pro` with search enabled. In (a) the request names the museum and asks about its status. The model finds out that the museum is closed and adjusts its recommendation. In (b) the request asks for an itinerary and never mentions the museum. The model recommends it as an afternoon stop and does not check whether it is open. The direct probe shows that the model can find the fact. The indirect probe evaluates whether it discovers the change on its own. The plot on the right shows the gap between direct and indirect probe accuracy.

[1] We will publicly release the benchmark and evaluation code upon publication.

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.

1 INTRODUCTION

AI agents, specifically LLMs equipped with search and browsing tools, are increasingly used for information seeking (Chatterji et al., 2025; Xi et al., 2025a). Access to real-time news articles and web pages through these tools invites the conclusion that agents' responses always reflect the current state of the world. We argue that this is not the case, making it all the more concerning that users often rely on these responses without verifying against the cited sources (Chapekis & Lieb, 2025).

Consider a user who asks an agent to plan their trip to Normandy (see Figure 1b). While the Bayeux Tapestry Museum would normally be an obvious recommendation, it is currently closed for renovations. The agent (Gemini-3.1-pro in the example in the Figure) fails to account for this real-world update and erroneously includes the museum in the user's itinerary. Crucially, such updates are easy for agents to find on the web. When directly asked to verify, the same agent readily informs the user that the museum is closed and suggests alternatives (Figure 1a). This inconsistency shifts the burden of verification to end users. Ideally, reliable agents must output well-researched and up-to-date responses by default. Agents must be able to identify unstated constraints (e.g. a museum's current status) and account for them, regardless of whether the user's query explicitly refers to them or asks to verify each fact. Without this, agents risk misleading users who trust their responses at face value.

Recent benchmarks, such as BrowseComp (Wei et al., 2025; Chen et al., 2026) or DeepSearchQA (Gupta et al., 2026), also investigate agentic search capabilities but on an orthogonal axis. They require complex long-horizon search but all the constraints that need to be satisfied are clearly specified. Similarly, benchmarks for temporal freshness (Vu et al., 2024; Meem et al., 2024; Kim et al., 2024) directly state what to search for or even include temporal markers (e.g., current) that suggest that facts may have changed. In our setting, on the other hand, relevant changes that impact the final answer are easy to find but the need to find them is not apparent from the query itself.

We introduce BLINDTOCHANGE, a framework to study whether search agents proactively discover recent events that affect their responses and account for their consequences. We measure this as the gap between agents' performance on two complementary probes of the same recent event: a direct probe that explicitly states the event in the query and an indirect probe that does not but requires reasoning about its consequences to answer correctly. This framing allows us to cleanly measure agents' ability to discover relevant events and reason about their impact, without confounding our analysis with the difficulty of finding the event itself.

BLINDTOCHANGE is constructed using an automated pipeline that extracts real-world events from Wikidata and leverages frontier LLMs to construct direct and indirect probes. Given a recent event, direct probes are easy to construct and responses to them are easy to evaluate. We phrase indirect probes as recommendation requests, carefully constructed such that the target event disqualifies an obvious candidate (e.g., the museum in our example). Then, we evaluate the agents' responses for the *absence* of the disqualified candidate. This allows us to objectively score long-form responses to a subjective query. We borrow this idea of constructing indirect probes as recommendation requests from Li & Goyal (2025). Their own framework is limited in scale as it was manually curated. Moreover, it supports evaluation only on a small retrieval corpus of synthetically generated events and data, which cannot match the complexity or scale of real-world news and internet data. Therefore, it is inadequate for evaluating frontier agentic search systems.

We make two key design decisions that ensure the validity of our evaluation framework. First, we restrict events to those that occurred after the reported knowledge cutoff of the agent. Second, we only retain pairs with indirect probes for which the agent included the disqualified candidate in their responses before the event. We approximate behavior before the event by sampling multiple outputs with search disabled. Together, these ensure that search is necessary to provide accurate answers for both direct and indirect probes, and that the *absence* of the disqualified candidate in a response is due to agentic search and not chance.

Our current dataset comprises direct and indirect probes based on two event types: demises and ends of official tenures of public figures. These were chosen as they have sufficient coverage in Wikidata in the months since the knowledge cutoff of the frontier models, although our framework extends to other event types. We evaluate 8 frontier models in two search harnesses, each model's own native search and a common scaffold with a shared web-search tool, at low and high reasoning effort. The native harness reflects how users actually encounter these systems, while the shared harness fixes the search backend across models and exposes the full search trajectory for analysis.

We find a large gap between what agents can find when asked explicitly and what they discover on their own. Agents almost always find the change when asked directly at high effort, reaching at least 96.5% direct accuracy on demises in nearly all configurations. Yet when queried with indirect probes for demises, the accuracy falls by a median 40.5 points at low effort and 27.5 points at high effort under the native harness; agents routinely recommend people who have passed away. The drop persists on tenures but is smaller, with median drops of 18.7 and 15.0 points for low and high reasoning effort, likely because officeholders change routinely and models are more inclined to check whether someone still holds a position. Strikingly, our error analysis under the shared harness finds that many failures stem from agents never verifying their own assumptions rather than from failing to retrieve the change. In fact, this accounts for over 75% of demise failures in 13 of the 15 configurations. Even on tenures, where models more often check whether the person still holds the position, not searching explicitly accounts for a median of 50% of failures across configurations.

One key finding is that strong performance on existing agentic benchmarks does not carry over to BLINDTOCHANGE. Notably, Claude-Opus-4.7 has the lowest indirect accuracy on demises under the shared harness at both effort levels. The issue is not search quality. When it searches as often as other models do, it is nearly as accurate. It simply searches less, issuing no search on 23% of demise probes at high effort (this number for other models is less than 8%). On tenures, where models search more readily, it is within 3 points of the most accurate model at both effort levels. GPT-5.4 shows the same pattern at low effort, trailing the open-weights models on demises by 34.0 to 52.5 points on a common set of probes as it issued no search for 58.5% of them. At high effort, it searches on 97.5% of demise probes and performs better than all other models.

Taken together, our work shows that finding information and recognizing when to look for it are distinct capabilities. Benchmarks that state what needs to be found therefore overestimate how well agents keep up with a changing world. Closing this gap requires systems that identify which of their own assumptions may be out of date and verify them without being asked.

2 RELATED WORK

This section summarizes notable related work. Extended discussion can be found in Appendix A.

Agentic Search Benchmarks. BrowseComp is a popular benchmark for agentic web search, with questions whose evidence is deliberately hard to locate, so answering requires searching both deeply and broadly (Wei et al., 2025). Other benchmarks directly score how the agent searches. DeepSearchQA (Gupta et al., 2026) measures whether agents find a complete set of answers and stops at the right time, InfoDeepSeek (Xi et al., 2025b) scores the quality of the information it gathers, and Mind2Web 2 (Gou et al., 2025) checks whether its answers are attributed to sources. LiveBrowseComp and LiveNewsBench separate search from recall by building questions from low-salience or recent facts that a model cannot have memorized (Fan et al., 2026; Zhang et al., 2026b).

Temporal Freshness Benchmarks. These tests models on questions whose answers change over time. RealTimeQA releases a new question set every week, drawn from news quizzes (Kasai et al., 2023). FreshQA has annotators re-verify answers every week to a fixed question set (Vu et al., 2024). To reduce manual effort, later benchmarks construct datapoints automatically from sources that record their own changes: PAT-Questions from time-dependent Wikidata relations (Meem et al., 2024), and EvolvingQA, GrowOVER, EvoWiki and DynaQuest from differences between successive Wikipedia snapshots (Kim et al., 2024; Ko et al., 2024; Tang et al., 2025; Lin et al., 2025).

However, both kinds of benchmarks explicitly state what information needs to be found, while we study whether agents discover information that the request never explicitly mentions.

Implicit Constraint Benchmarks. Task success often depends on factors the user never explicitly states but expects an assistant to recognize (Kaur et al., 2026). Several benchmarks make such factors discoverable through tool use. Implicit Intelligence hides constraints in the environment (Sirdeshmukh & Wetter, 2026), DeepPlanning checks travel and shopping plans against implicit constraints such as limited seat availability (Zhang et al., 2026a), and MapSatisfyBench recovers the decision factors users left unspoken from behavior logs (Bai et al., 2026). All three find that agents rarely probe for the missing information, even when a tool call would reveal it. We show that implicit constraints also arise when seeking information about a changing world, and find that agents similarly fail to search for and discover them using web search.

3 BLINDTOCHANGE

We have the following desiderata for BLINDTOCHANGE: (i) It must measure whether agents proactively discover and account for updates that affect the correctness of their answers; (ii) its evaluation must be objective and reliable; and (iii) it must be automatically constructible, and hence easy to scale and refresh. We first describe the conceptual ideas and design choices that help realize these desiderata in §3.1. We then describe our automated pipeline for constructing the benchmark in §3.2.

3.1 CONCEPTUAL FRAMEWORK

Setup. Consider an event that changes a fact about an entity x from f_{old} to f_{new}. In Figure 1, x is the *Bayeux Tapestry Museum* and the event is its closure. The fact f_{old} is that the museum is open to visitors, and f_{new} is that it is closed until October 2027. While the event likely changes other facts too, we focus only on the update $u = (f_{\text{old}}, f_{\text{new}})$.

We can represent the knowledge state of the world before and after the event as K_{old} and K_{new} respectively. Both knowledge states contain a combination of common facts that are unchanged by the event, as well as facts that are outdated (in K_{old}) or updated (in K_{new}).

For a given question q, let $\mathcal{A}(q; K)$ denote the set of valid answers under knowledge state K. We use a set because many questions, such as requests for recommendations, admit several valid answers. We then define the *probes for an update* u as the set of questions for which the valid answers based on knowledge before and after the update differ:

$$\mathcal{Q}(u) = \{q : \mathcal{A}(q; K_{\text{old}}) \neq \mathcal{A}(q; K_{\text{new}})\}.$$

We want to measure whether an agent's response reflects the updated knowledge K_{new} when queried with probes in $\mathcal{Q}(u)$. We restrict probes to events that occur after the knowledge cutoff of the agent being evaluated, so that the agent's parametric knowledge does not contain f_{new}. At the same time, the agent always has access to K_{new} through real-time search. We can therefore attribute a response that correctly reflects K_{new} to successful agentic search. Our setup involves several key choices:

Design Decision 1: Paired direct and indirect probes for evaluation. Recall that we want to measure whether an agent will proactively search for consequential updates or events when q does not explicitly mention them. However, we do not want to confound our analysis with the difficulty of finding the update itself. To separate the two, we define two types of complementary probes that operate on the same underlying update and compare performance on them.

Let $\mathcal{Q}_{\text{dir}}(u) \subseteq \mathcal{Q}(u)$ be questions that explicitly ask which of f_{old} or f_{new} holds. We call these ***direct probes***. For example, "*Will the Bayeux Tapestry Museum be open to visitors?*" in Figure 1a is a direct probe. For these questions, no additional reasoning or inference on top of the updated fact f_{new} is needed to answer the question correctly. Merely knowing the fact suffices. Therefore, these direct probes serve as a *control* that allows us to isolate the difficulty of finding the update itself, given access to K_{new}. As they explicitly inquire about the update, answers to direct probes are straightforward to verify with an LLM judge.

We define ***indirect probes*** as all probes for u that are not direct probes:

$$\mathcal{Q}_{\text{ind}}(u) = \mathcal{Q}(u) \setminus \mathcal{Q}_{\text{dir}}(u).$$

The question *"Suggest a five-day itinerary in Normandy"* in Figure 1b is an example of an indirect probe. Although it never mentions the museum or asks about its status, the update changes its set of valid answers. An itinerary that includes the *Bayeux Tapestry Museum* is no longer valid. Our benchmark reports each agents' performance on both direct and indirect probes.

There are two main challenges in constructing indirect probes. (i) The question must not explicitly inquire about f_{new} or the event, yet the update must still change the set of valid answers. (ii) An agent may avoid an invalid answer by chance rather than by discovering the update. For example, it may omit the *Bayeux Tapestry Museum* from an itinerary without ever learning of the closure. We must distinguish such cases from genuine discovery. Our design and selection of indirect probes address both challenges.

Design Decision 2: Instantiating indirect probes as recommendation requests. Our indirect probes are recommendation requests for which the entity x is a natural candidate before the update but is disqualified after it. Formally, let $\mathcal{C}(q; K)$ denote the set of entities eligible for the request q under knowledge state K. Indirect probes, other than being in $\mathcal{Q}_{\text{ind}}(u)$, must satisfy:

$$x \in \mathcal{C}(q; K_{\text{old}}) \text{ and } x \notin \mathcal{C}(q; K_{\text{new}}).$$

In other words, the update makes x ineligible for the request. Li & Goyal (2025) introduce the term indirect probes for a specific construction, namely questions with list-style answers that include x before the update but not after it. We instead give a formal definition of indirect probes as $\mathcal{Q}_{\text{ind}}(u)$ and adopt a similar construction to instantiate our benchmark.

Note that these recommendation-style indirect probes address challenge (i) above, but not challenge (ii) which we address next via design decision 3.

Design Decision 3: Selecting indirect probes where the agent consistently recommends x before the update. Suppose the agent being evaluated consistently recommends x for an indirect probe before the update, relying on f_{old}. We can then score their answers under K_{new} (our actual evaluation setting) by the *absence* of x as a recommendation. E.g., if the agent consistently recommended *Bayeux Tapestry Museum* before the closure, then its omission after the update can be reasonably attributed to having discovered the update.

To determine whether an agent consistently recommends x under K_{old}, we would need to block its access to every web page that reports the update. This is infeasible. However, we choose events that occur after the agent's knowledge cutoff, we therefore can sample responses from the agent *with search disabled* and use them as a proxy for what it would have recommended before the update.

For each candidate indirect probe, we sample multiple responses with search disabled and retain the probe only if the agent recommends x in every response. This strict criterion addresses the (ii) concern, and allows us to attribute the omission of x — when search is enabled — to discovering the update rather than to sampling variation.

3.2 Automatic Benchmark Construction Pipeline

Each datapoint in our benchmark is a pair of direct and a complementary indirect probe that target the same underlying update. We next describe how we gather these updates and construct the probes automatically. Appendix B gives further details on each step.

Gathering updates. We source real-world events from Wikidata. We focus on two event types about people, namely demises and ends of official tenures (tenures for short). We choose these as proofs of concept because both have enough support in Wikidata since frontier models' knowledge cutoffs, although our conceptual framework applies to other event types.

Constructing indirect probes. For each update, we prompt a frontier LLM to generate direct and indirect probes. Design decision 3 retains an indirect probe for an agent only if the agent consistently recommends x without search. We therefore first generate and filter indirect probes, and then generate direct probes only for the filtered subset. We provide the update and associated Wikipedia articles as context to the LLM and prompt it to propose five candidate indirect probes. Each probe asks for a shortlist that admits several plausible candidate entities without uniquely identifying the entity x that the update is about. We also include category-specific constraints. For probes related to demises, the request must call for future participation. For tenures, the request must require authority that only the current holder of the position has. Across the five probes for an update, we vary who is making the request (e.g., a doctoral student), the scenario it arises in (e.g., seeking approval for a research visit), and which consequence of the update makes x ineligible (only the department head can grant that approval, so if their tenure is over, they would be ineligible).

Filtering indirect probes per agent. Following design decision 3, we sample three responses with search disabled for each agent and candidate indirect probe and retain the probe only if all three recommend x. An LLM judge decides whether each response recommends x. It is carefully prompted so that it only penalizes a response if it *recommends* the disqualified entity, and does not penalize mere mentions. We filter probes separately for each agent, since agents differ in what they recommend. We filter at low reasoning effort, since what a model knows depends on its weights rather than its reasoning budget, and share the resulting subset across both effort settings so that their comparison uses the same questions.

Final benchmark and scoring. For each retained indirect probe, we prompt the LLM to generate a paired direct probe that explicitly asks which of f_{old} and f_{new} holds. An LLM judge scores a direct response as correct only if it explicitly states f_{new}, namely that the person has died (demises) or

no longer holds the position (tenures). Responses that affirm f_{old}, express uncertainty, or evade the question are marked incorrect. For indirect probes, we reuse the judge from the filtering step.

Implementation Details. We generate indirect probes with Gemini-3.1-Pro (Google DeepMind, 2026a) and direct probes with Gemini-3.7-Flash (Google DeepMind, 2026c). We use Gemini-3.5-Flash-Lite (Google DeepMind, 2026b) as the judge for all indirect and direct responses, including those scored during filtering. Appendix B lists the reasoning settings for each stage, and the full prompts are provided in the supplementary material. To validate the judge, one author labeled a stratified sample of 200 responses per category for each probe type, without access to the judge's labels. On indirect probes, the judge agreed with the human label on 199 (99.5%) responses for demises and 197 (98.5%) for tenures. On direct probes, it agreed on all 200 responses for tenures and on 198 (99.0%) for demises. See Appendix C for more details.

Benchmark Statistics. We instantiate our benchmark with events dated between Feb and June 2026, yielding 2,078 demises and 751 tenures. The window begins in Feb because the latest reported cutoff among the models we evaluate is January 2026, that of Claude-Opus-4.7 (Table 5). Generating five indirect probes per event gives 10,390 candidate probes for demises and 3,755 for tenures. Table 6 reports the fraction of retained probes for each model after the filtering step described above.

4 Experimental Setup

Models Evaluated. We evaluate 8 frontier LLMs from different providers, covering both open- and closed-source models: Claude-Opus-4.7 (Anthropic, 2026a), GPT-5.4 (OpenAI, 2026a), Gemini-3.1-Pro (Google DeepMind, 2026a), Qwen3.7-Max (Qwen Team, 2026), DeepSeek-V4-Pro-0423 (Xu et al., 2026), Kimi-K2.6 Kimi Team (2026), MiMo-V2.5-Pro (Xiaomi, 2026a), and GLM-5.1 (GLM Team, 2026). Providers expose reasoning effort in different ways, so we evaluate each model at two settings, chosen as the nearest available match to a low and a high effort budget. We refer to these as low and high throughout, and Table 7 lists the provider-specific values.

From each model's probes retained by the filtering step (Table 6) we sample 200 per category, drawing exactly one probe per event. Where fewer than 200 events survive the filter, we use all of them. This affects tenures for Claude-Opus-4.7, with 175 events, and MiMo-V2.5-Pro, with 115. All results in Table 1 are computed on these model-specific samples.

Native and shared harnesses for evaluation. Most providers ship a *native* harness with a first-party search tool, the provider's default system instructions, and other specifications. These reflect how users actually encounter these systems. However, these harnesses differ across providers and offer a limited view of the search trajectories, hindering both apples-to-apples comparison and systematic analysis. Therefore, we additionally evaluate every model under a common scaffold, which we call the *shared* harness. Models without a *native* harness are only evaluated using the *shared* harness.

We implement the *shared* harness with Inspect AI (UK AI Security Institute, 2024). We fix the search backend and interaction protocol so that differences between models are not due to provider-specific search integrations. Each model receives the question and a common system prompt (Figure 4) that states the evaluation date, the model's knowledge cutoff, and instructions to search before answering potentially time-sensitive questions. The harness exposes a single web-search tool backed by Exa (`https://exa.ai`), which returns source titles, URLs, publication dates when available, and relevant text highlights. Models write their own queries and may search repeatedly before answering. We log every response, search call and tool output for post-hoc analysis.

5 Main Results: The Gap Between Direct and Indirect Probing

Table 1 reports results for all 8 models. We refer to a model, harness and reasoning effort together as a configuration. For each configuration, we report accuracy under direct and indirect probes in the Dir. and Ind. columns respectively. The ▼ columns report the gap between the two.

Indirect vs direct probing. Indirect accuracy falls below direct accuracy in every configuration, with one exception (Qwen3.7-Max with the native harness). On demises, the median gap is 42.5 points at low effort and 21.8 points at high effort with the shared harness. With the native harness that users actually encounter, the gap is large at 40.5 and 27.5 points respectively. Notably, Claude-Opus-4.7, a model at the frontier of other agentic search benchmarks, shows the largest gap at 70.0 points with the shared harness at low effort. Overall, results in Table 1 show that models that can

Table 1: Direct (Dir.) and indirect (Ind.) probing accuracy (%) for each model, harness and reasoning effort. The ▼ columns show the drop from direct to indirect accuracy. – marks MiMo-V2.5-Pro (shared, low), which kept calling tools without giving a final answer.

		Demises						Tenures					
		Low			High			Low			High		
Model	Harness	Dir.	Ind.	▼	Dir.	Ind.	▼	Dir.	Ind.	▼	Dir.	Ind.	▼
Opus-4.7	native	96.0	58.5	37.5	96.5	69.0	27.5	97.7	84.6	13.1	96.0	85.7	10.3
	shared	96.5	26.5	70.0	97.0	60.0	37.0	96.0	90.9	5.1	100.0	91.4	8.6
5.4	native	99.0	57.5	41.5	100.0	92.0	8.0	86.0	76.5	9.5	98.5	93.5	5.0
	shared	99.5	35.5	64.0	100.0	94.5	5.5	96.5	83.5	13.0	99.5	94.0	5.5
3.1-Pro	native	99.5	54.0	45.5	100.0	87.5	12.5	99.5	60.0	39.5	99.5	85.5	14.0
	shared	99.5	44.0	55.5	99.5	75.5	24.0	99.0	74.5	24.5	100.0	83.5	16.5
3.7-Max	native	21.0	30.0	-9.0	100.0	72.5	27.5	70.5	34.0	36.5	99.0	83.0	16.0
	shared	73.5	31.0	42.5	100.0	77.5	22.5	95.5	48.5	47.0	100.0	86.0	14.0
V4-Pro	shared	97.5	77.5	20.0	100.0	84.0	16.0	92.5	91.5	1.0	100.0	93.0	7.0
K2.6	shared	98.5	78.5	20.0	100.0	78.5	21.5	95.0	81.5	13.5	100.0	87.0	13.0
V2.5-Pro	native	85.0	32.5	52.5	86.0	34.5	51.5	86.1	68.7	17.4	89.6	72.2	17.4
	shared	93.0	–		100.0	78.0	22.0	99.1	–		100.0	92.2	7.8
5.1	native	98.0	58.5	39.5	98.5	64.5	34.0	84.5	64.5	20.0	91.0	74.5	16.5
	shared	94.5	66.5	28.0	100.0	86.5	13.5	99.5	92.5	7.0	99.5	93.5	6.0

retrieve an update when asked about it often fail to discover and account for the update when the request does not explicitly inquire about it.

Demises vs tenures. Interestingly, we find that the gap is typically larger on demises than on tenures (in 23/27 configurations). With the native harness, the median gap on tenures is 18.7 points compared to 40.5 on demises at low effort, and 15.0 compared to 27.5 at high effort. The shared harness shows the same pattern, with 13.0 compared to 42.5 points at low effort and 8.2 compared to 21.8 at high effort. Gemini-3.1-Pro and Qwen3.7-Max stand out at low effort, with tenure gaps of 39.5 and 36.5 points under the native harness. We hypothesize that because officeholders change routinely, models are more inclined to check whether a person still holds a position than whether they are still alive. We examine this in Section 6.

High vs low reasoning effort. Higher effort improves indirect accuracy on demises in 12/13 model-harness pairs, with a median gain of 31.5 points. However, the benefit is uneven. With the shared harness, GPT-5.4 improves from 35.5% to 94.5%, while Kimi-K2.6 stays at 78.5%. On tenures, indirect accuracy improves for all 13 pairs, but by a smaller median 9.0 points. We hypothesize that higher effort helps mainly by prompting models to check whether the recommended person is still alive or still holds the position. Models already do this more often for positions at low effort, which leaves less room to improve on tenures. Section 6 confirms this.

Claude-Opus-4.7 rarely searches on demise probes. Although Claude-Opus-4.7 performs strongly on tenures and on other popular agentic benchmarks (Anthropic, 2026a), it has the largest gap on demises with the shared harness at both effort levels (70.0 and 37.0 points). Figure 2 groups the demise probes of Claude-Opus-4.7 and three other models by the number of search queries issued (see Appendix H for the remaining models). Among probes with at least two searches, Claude-Opus-4.7 answers 79% correctly at low effort and 78% at high effort, close to the median of 82% across the other shared configurations in Table 1. However, it issues no search on 68% of demise probes at low effort and 23% at high effort, while no other model exceeds 8% at high effort. Its larger gap therefore reflects how rarely it searches, not how well it uses what it finds.

Shared vs native harness. Our analysis of search trajectories relies on the shared harness, so models should perform comparably in both harnesses. At high effort they do, with indirect accuracy higher in the shared harness for 9 of the 12 model-category pairs and at most 12.0 points lower for the rest. At low effort, the shared harness trails on demises for Claude-Opus-4.7, GPT-5.4 and Gemini-3.1-Pro by up to 32.0 points, possibly because with reduced or no reasoning these models use a generic search tool less effectively than their provider's own. We therefore consider the shared harness a reasonable setting in which to study search behavior.

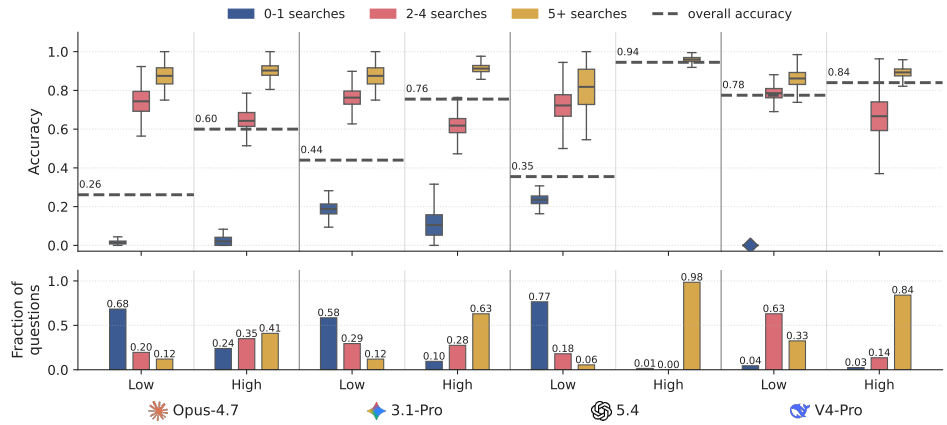


Figure 2: Across models, we observe comparable performance on indirect probes for demises using the shared harness, when grouped by the number of search queries issued. Top, accuracy per group (with at least 9 probes) and its overall accuracy (dashed). Bottom, the share of probes in each group.

Comparing models. To compare models on equal footing, we evaluate five models under the shared harness on a common set of probes retained by all of them (Table 2). We include these five models to retain a sufficient number of probes in the intersection for a reliable comparison. At low effort, the three open-weights models are more accurate than the two closed ones on both categories. Search behavior accounts for much of this difference. GPT-5.4 and Gemini-3.1-Pro issue no search on 58.5% and 56.5% of demise probes, compared with at most 30.0% for the open-weights models. When Gemini-3.1-Pro does search, it answers 80.5% of probes correctly, in line with the open-weights models. On tenures, the closed models rarely skip search but search less, issuing a median of 3 queries against 7 to 11 for the open-weights models. At high effort, GPT-5.4 is the most accurate model on demises and performs comparably to the open-weights models on tenures. It issues a median of 16 queries on demise probes, more than any other model, and searches on 97.5% of them. The spread across models is also larger on demises, at 52.5 points at low effort and 18.5 points at high effort, than on tenures, at 12.8 and 7.5 points. This is likely because models search more readily on tenures, skipping search less often than on demises in all configurations.

Table 2: Indirect probing accuracy (%) on a common set of questions shared across all models, run using the shared harness with both low and high reasoning effort.

	Demises (n = 200)		Tenures (n = 172)	
Model	Low	High	Low	High
5.4	26.0	90.5	76.7	89.5
3.1-Pro	35.5	72.0	79.1	82.6
V4-Pro	65.0	81.5	89.5	90.1
K2.6	78.5	79.0	83.1	87.8
5.1	60.0	82.5	83.7	88.4

6 Error Analysis

To understand the gap between direct and indirect accuracy, we further categorize failures on indirect probes. We perform this analysis under the shared harness as it logs every query, every search result and the context the model has before answering. We hypothesize that a model can fail at three stages: it never looks for the update, it looks for it but retrieval fails, or retrieves it and still gives an outdated recommendation. We convert these stages into four mutually exclusive failure modes. We assign each failure to the first one it satisfies, in the order below.

(1) **Reasoning error.** The model's context contains evidence of the update, yet its answer still makes the invalid recommendation. Prior work finds that models can favor parametric memory over conflicting context (Appendix A). (2) **Ineffective but explicit search.** The context contains no conclusive evidence of the update, but at least one query directly asks whether f_{old} still holds, such as whether the person is alive or who currently holds the position. Such failures often reflect poorly phrased queries rather than limitations of the search tool. (3) **Indirect search.** No query asks about the update explicitly, but at least one could reveal it indirectly, such as a search for the person's recent public appearances or actions in the role. Because our events are recent, such searches often

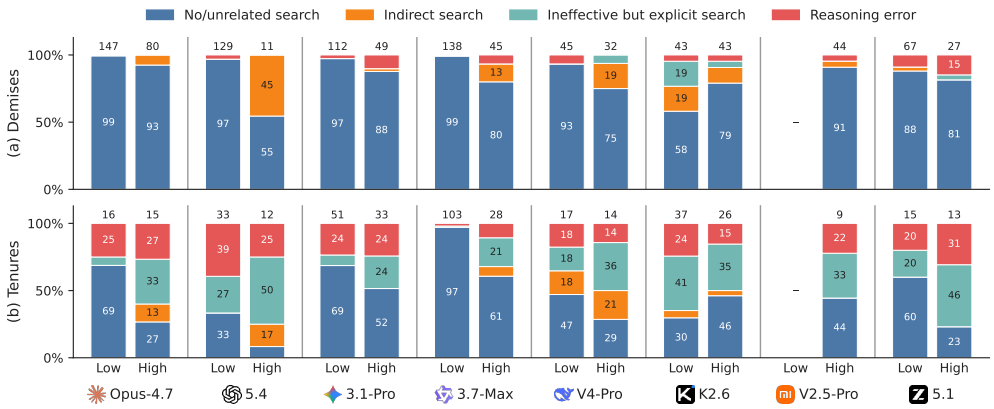


Figure 3: Composition of indirect-probe failure modes under the shared harness for (a) demises and (b) tenures. Numbers above bars give the total number of failures.

return activity from before the update, which can suggest that f_{old} still holds. (4) **No/unrelated search.** No query could reveal the update, either because the model does not search or because its searches concern other parts of the request.

We assign failure categories to trajectories with Gemini-3.7-Flash at low reasoning effort (prompt is in the supplementary material). We verified the assigned mode for a stratified sample of 160 demise and 165 tenure failures, with an agreement of 98.1% and 81.2% respectively (Appendix I). Most tenure disagreements involve searches whose relevance depends on background knowledge of the position, such as a search for the current members of a Senate committee, which would omit a senator who has left office. These mainly move failures between indirect and unrelated search, so our tenure conclusions rest on the two modes together. Appendix J shows representative failures.

Not searching explicitly drives the gap, especially on demises. On demises, failures with no/unrelated search make up at least 75% of failures in 13 of the 15 configurations (Figure 3a). On tenures, failures are spread more evenly across the four modes (Figure 3b). Even so, indirect and no/unrelated search together account for at least 23% of failures in every configuration and for a median of 50% of failures. In 14 of the 15 configurations, failures with no/unrelated search account for all of the additional failures on demises compared with tenures. Since direct accuracy is similar on both categories, this explains the larger gap on demises.

Higher reasoning helps mainly by getting models to search for the update. We compare the share of probes in each failure mode at low and high effort for the seven models evaluated at both effort levels under the shared harness. On demises, moving from low to high effort reduces failures with no/unrelated search by 28.9 points on average, nearly matching the average gain in indirect accuracy of 28.1 points. On tenures, the reduction is 9.1 points and the gain is 9.4 points. No other failure mode changes by more than 5.0 points for any model. Higher effort therefore improves accuracy almost entirely by getting models to search for the update on probes where they previously did not. The gain is larger on demises because failures with no/unrelated search are far more common there at low effort, which leaves more room to improve. Kimi-K2.6 is the exception and does not improve on demises because it already has few failures with no/unrelated search at low effort.

7 Conclusion

We introduce BlindToChange to track a new axis of agentic search capabilities: do AI agents proactively discover recent real-world changes that affect their responses? By pairing each indirect probe with a direct probe of the same change, we separate an agent's ability to find a change when asked from its ability to discover the change unprompted and account for its consequences. We show that frontier agents almost always find a change when asked directly, yet often miss it when the request does not inquire about it. Our error analysis traces many of these failures to agents never explicitly checking whether the relevant fact has changed, and shows that higher reasoning effort helps mainly by prompting such checks. Our findings highlight that benchmarks which state what needs to be found overestimate how well agents keep up with a changing world.

AI USE STATEMENT

In this work, we used generative AI tools to generate the direct and indirect probes that make up our benchmark (§3.2), to judge model responses, and to support qualitative analysis by labeling failure modes in our error analysis (§6). We validated both the judge and the failure-mode labels against human annotations (Appendices C and I). We also used AI coding assistants to help implement parts of our pipeline. We have not used generative AI tools to propose or refine hypotheses, design or provide feedback on our methodology or experiments, clean or reformat the dataset, interpret results, or develop our conceptual framework. Formulating mathematical claims, providing ingredients for or writing proofs, and translation are not applicable to this work. Additionally, we used generative AI tools to edit the paper for readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anthropic. System Card: Claude Opus 4.7. Technical report, Anthropic, April 2026a. URL https://www.anthropic.com/claude-opus-4-7-system-card. Accessed: 2026-09-10.

Anthropic. Anthropic transparency hub. https://www.anthropic.com/transparency, 2026b. Accessed: 2026-09-10.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Lubin Bai, Mengyu Cao, Sixue Wang, Zhongwei Wan, Yue Pan, Jiale Hou, Xiang Li, and Xiuyuan Zhang. Mapsatisfybench: Benchmarking satisfaction-aware map agents through behavior-grounded implicit decision factors. *arXiv preprint arXiv:2606.17453*, 2026.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. Stale: Can llm agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Athena Chapekis and Anna Lieb. Google users are less likely to click on links when an AI summary appears in the results. Pew Research Center, July 2025. URL https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research, September 2025. URL http://www.nber.org/papers/w34255.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-plus: A fair and disentangled evaluation benchmark for deep search agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22349–22370, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1023. URL https://aclanthology.org/2026.acl-long.1023/.

Jiehan Cheng and Zhicheng Dou. Dailyqa: A benchmark to evaluate web retrieval augmented llms based on capturing real-world changes. *arXiv preprint arXiv:2505.17162*, 2025.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Mingyi Deng, Lijun Huang, Yani Fan, Fanqi Kong, Jiayi Zhang, Fashen Ren, Jinyi Bai, Fuzhen Yang, Dayi Miao, Zhaoyang Yu, Yifan Wu, Yanfei Zhang, Fengwei Teng, Yingjia Wan, Song Hu, Yude Li, Xin Jin, Conghao Hu, Haoyu Li, Qirui Fu, Tai Zhong, Xinyu Wang, Xiangru Tang, Nan Tang, Chenglin Wu, and Yuyu Luo. Interactcomp: Evaluating search agents with ambiguous queries. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AiawINF2T0.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. Context versus prior knowledge in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13211–13235, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.714. URL https://aclanthology.org/2024.acl-long.714/.

HuiMing Fan, Xiao Wang, Zheng Chu, Qianyu Wang, Zhuoyao Wang, Ming Liu, Bing Qin, et al. Livebrowsecomp: Are search agents searching, or just verifying what they already know? *arXiv preprint arXiv:2605.28721*, 2026.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. Tracking the limits of knowledge propagation: How LLMs fail at multi-step reasoning with conflicting knowledge. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5813–5847, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.273. URL https://aclanthology.org/2026.eacl-long.273/.

Constanza Fierro, Nicolas Garneau, Emanuele Bugliarello, Yova Kementchedjhieva, and Anders Søgaard. MuLan: A study of fact mutability in language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 762–771, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.67. URL https://aclanthology.org/2024.naacl-short.67/.

GLM Team. GLM-5.1: Towards Long-Horizon Tasks. Technical report, Z.ai, April 2026. URL https://z.ai/blog/glm-5.1. Accessed: 2026-09-10.

Google. Gemini 3 Developer Guide. https://ai.google.dev/gemini-api/docs/gemini-3, 2026. Accessed: 2026-09-20.

Google DeepMind. Gemini 3.1 Pro Model Card. Technical report, Google DeepMind, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-09-10.

Google DeepMind. Gemini 3.5 Flash-Lite Model Card. Technical report, Google DeepMind, July 2026b. URL https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/. Accessed: 2026-09-10.

Google DeepMind. Gemini 3.7 Flash Model Card. Technical report, Google DeepMind, August 2026c. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/. Accessed: 2026-09-10.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, 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 Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/202

5/hash/fdcec9f5b99aa4fc8f4fb8487802d737-Abstract-Datasets_and_Benchmarks_Track.html.

Nikita Gupta, Riju Chatterjee, Lukas Haas, Connie Tao, Andrew Wang, Chang Liu, Hidekazu Oiwa, Elena Gribovskaya, Jan Ackermann, John Blitzer, et al. Deepsearchqa: Bridging the comprehensiveness gap for deep research agents. *arXiv preprint arXiv:2601.20975*, 2026.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran T. Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c63819755591ea972f8570beffca6b1b-Abstract-Datasets_and_Benchmarks_Track.html.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Nimit Kalra, Narutatsu Ri, Zerzar Bukhari, Ang Li, Sanae Lotfi, Liam H Fowl, and Micah Goldblum. Measuring the limits of continual learning for LLMs. In *ICML 2026 Workshop on Combining Theory and Benchmarks: Towards A Virtuous Cycle to Understand and Guarantee Foundation Model Performance*, 2026. URL https://openreview.net/forum?id=BlDEHKSGy9.

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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb7-Abstract-Datasets_and_Benchmarks.html.

Kirandeep Kaur, Vinayak Gupta, Aditya Gupta, and Chirag Shah. PROPER agents: Proactivity driven personalized agents for advancing knowledge gap navigation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41951–41975, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2082. URL https://aclanthology.org/2026.findings-acl.2082/.

Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. Carpe diem: On the evaluation of world knowledge in lifelong language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5401–5415, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.302. URL https://aclanthology.org/2024.naacl-long.302/.

Kimi Team. Kimi K2.6: Advancing Open-Source Coding. Technical report, Moonshot AI, April 2026. URL https://www.kimi.ai/blog/kimi-k2-6. Accessed: 2026-09-10.

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. GrowOVER: How can LLMs adapt to growing real-world knowledge? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3282–3308, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.181. URL https://aclanthology.org/2024.acl-long.181/.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. Studying large language model behaviors under context-memory conflicts with real documents. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=xm8zYRfrqE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Aochong Oliver Li and Tanya Goyal. Memorization vs. reasoning: Updating LLMs with new knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25853–25874, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1326. URL https://aclanthology.org/2025.findings-acl.1326/.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, et al. Horizonbench: Long-horizon personalization with evolving preferences. *arXiv preprint arXiv:2604.17283*, 2026.

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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.276. URL https://aclanthology.org/2025.emnlp-main.276/.

Qian Lin, Junyi Li, and Hwee Tou Ng. DynaQuest: A dynamic question answering dataset reflecting real-world knowledge updates. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26918–26936, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1380. URL https://aclanthology.org/2025.findings-acl.1380/.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Jannat Meem, Muhammad Rashid, Yue Dong, and Vagelis Hristidis. PAT-questions: A self-updating benchmark for present-anchored temporal question-answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13129–13148, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.777. URL https://aclanthology.org/2024.findings-acl.777/.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. GPT-5.4 Thinking System Card. Technical report, OpenAI, March 2026a. URL https://openai.com/index/gpt-5-4-thinking-system-card/. Accessed: 2026-09-10.

OpenAI. GPT-5.4. https://developers.openai.com/api/docs/models/gpt-5.4, 2026b. Accessed: 2026-09-20.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6036–6063, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.301. URL https://aclanthology.org/2025.acl-long.301/.

Bhawna Piryani, Zehra Mert, and Adam Jatowt. How often do answers change? estimating recency requirements in question answering. *arXiv preprint arXiv:2603.16544*, 2026.

Sergey Pletenev, Maria Marina, Nikolay Ivanov, Daria Galimzianova, Nikita Krayko, Mikhail Salnikov, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. Will it still be true tomorrow? multilingual evergreen question classification to improve trustworthy QA. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8603–8620, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.434. URL https://aclanthology.org/2025.emnlp-main.434/.

Qwen Team. Qwen3.7: The Agent Frontier. Technical report, Alibaba Group, May 2026. URL https://qwen.ai/blog?id=qwen3.7. Accessed: 2026-09-10.

Ved Sirdeshmukh and Marc Wetter. Implicit intelligence–evaluating agents on what users don't say. *arXiv preprint arXiv:2602.20424*, 2026.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, and Yong Liao. EvoWiki: Evaluating LLMs on evolving knowledge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 948–964, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.47. URL https://aclanthology.org/2025.acl-long.47/.

Yiling Tao, Shihan Deng, Meiling Tao, Pengzhi Wei, Zhichao Hu, and Zhihao Zhu. When search agents should ask: Discobench for clarification-aware deep search. *arXiv preprint arXiv:2606.27669*, 2026.

UK AI Security Institute. Inspect AI: Framework for large language model evaluations. https://github.com/UKGovernmentBEIS/inspect_ai, 2024. Accessed: 2026-09-10.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Jonas Wallat, Wolfgang Nejdl, and Sandipan Sikdar. When facts change: Temporal knowledge conflict resolution in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2154–2184, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.103. URL https://aclanthology.org/2026.findings-acl.103/.

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.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*, 2025a.

Yunjia Xi, Jianghao Lin, Menghui Zhu, Yongzhao Xiao, Zhuoying Ou, Jiaqi Liu, Tong Wan, Bo Chen, Weiwen Liu, Yasheng Wang, et al. Infodeepseek: Benchmarking agentic information seeking for retrieval-augmented generation. *arXiv preprint arXiv:2505.15872*, 2025b.

Xiaomi. MiMo-V2.5-Pro. Technical report, Xiaomi, April 2026a. URL https://mimo.xiaomi.com/mimo-v2-5-pro/. Accessed: 2026-09-10.

Xiaomi. MiMo deep thinking usage guide. https://mimo.mi.com/docs/en-US/quick-start/usage-guide/text-generation/deep-thinking, 2026b. Accessed: 2026-09-20.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Yinger Zhang, Shutong Jiang, Renhao Li, Jianhong Tu, Yang Su, Lianghao Deng, Xudong Guo, ChenXu Lv, and Junyang Lin. DeepPlanning: Benchmarking long-horizon agentic planning with verifiable constraints. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7377–7407, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.335. URL https://aclanthology.org/2026.acl-long.335/.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. Livenewsbench: Evaluating llm web search capabilities with freshly curated news. *arXiv preprint arXiv:2602.13543*, 2026b.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A Extended Related Work

Agentic Search Systems and Benchmarks. Traditional Retrieval-Augmented Generation (RAG) systems supplement a model's parametric memory with a non-parametric retrieval corpus that is queried once using the input question (Lewis et al., 2020). A line of work later made the retrieval decision part of generation itself. FLARE anticipates the upcoming sentence and, when that anticipation is low confidence, uses it to form a retrieval query (Jiang et al., 2023). Self-RAG replaces this heuristic with learned reflection tokens, so the model decides at each step whether to retrieve and whether the retrieved evidence supports what it wrote (Asai et al., 2024). A parallel line of work gives the model a set of actions, now commonly called tool calls, rather than a single retrieve operation. WebGPT lets a model search, click, scroll, and quote inside a live browser (Nakano et al., 2021), and ReAct interleaves reasoning with acting, so the model can revise its plan from what it observes (Yao et al., 2023). The agentic search systems we evaluate, often called agentic RAG, combine the two lineages (Li et al., 2025). The model chooses when to invoke a search or browse tool, what query or URL to invoke it with, and when it has gathered enough evidence to answer.

Benchmarks for these systems have followed. GAIA poses general assistant tasks that require tool use, browsing, and multimodal reasoning (Mialon et al., 2024). BrowseComp narrows the setting to web research, with short-answer questions whose evidence is deliberately entangled and hard to locate (Wei et al., 2025). Because the live web changes between runs, BrowseComp-Plus rebuilds the same style of evaluation over a fixed human-verified corpus, which separates retrieval quality from reasoning over what was retrieved (Chen et al., 2026). Other benchmarks shift the evaluation object from the final answer to the search process itself. DeepSearchQA scores comprehensiveness and reasoned stopping on exhaustive-list web tasks (Gupta et al., 2026), InfoDeepSeek scores the accuracy, utility, and compactness of the information an agent gathers from the live web (Xi et al., 2025b), and Mind2Web 2 judges long-horizon trajectories for both answer correctness and source attribution (Gou et al., 2025). Two recent benchmarks ask how much of this performance is search rather than recall. LiveBrowseComp builds questions from low-salience facts published in the preceding ninety days, and shows that they largely defeat agents that score well on BrowseComp (Fan et al., 2026). LiveNewsBench generates its questions automatically from clusters of news articles about a single recent event and refreshes them quarterly, so both the questions and answers change (Zhang et al., 2026b).

Temporal Freshness Benchmarks. Facts differ in how long they stay true. MuLan separates Wikidata relations into mutable and immutable classes and finds that closed-book models are both less accurate and markedly less confident on the mutable ones (Fierro et al., 2024). EverGreenQA labels questions rather than facts, and finds that this label predicts retrieval behavior (Pletenev et al., 2025). RecencyQA also classifies how often an answer changes, from within an hour to never, and notes that this rate can itself depend on context (Piryani et al., 2026).

Benchmarks have taken up how well models handle questions whose answers change. RealTimeQA draws its questions from weekly news quizzes and releases a new set every week (Kasai et al., 2023). FreshQA instead fixes a set of 600 questions written by annotators to read like search queries, grouped by how quickly their answers turn over, and its maintainers re-verify the answers every week (Vu et al., 2024). Keeping either benchmark current is manual work, and its answers can fall out of date between updates. Later benchmarks instead construct questions from a source that records its own changes, so that answers can be refreshed automatically. PAT-Questions builds questions from templates over time-dependent Wikidata relations, and refreshes every answer by re-running the SPARQL query that produced it (Meem et al., 2024). EvolvingQA, GrowOVER, EvoWiki, and DynaQuest instead compare successive Wikipedia snapshots and generate questions from the passages or infobox fields that differ between versions (Kim et al., 2024; Ko et al., 2024; Tang et al., 2025; Lin et al., 2025). DailyQA follows the same recipe on a daily cycle, using revision logs to keep the answer key in step (Cheng & Dou, 2025).

Unstated Constraints and Agent-Initiated Information Seeking. Kaur et al. (2026) represent the factors that determine task success as dimensions and separate the ones a user has stated from the ones they have not, treating the unstated ones as gaps an assistant should recognize on the user's behalf. Benchmarks that make such gaps discoverable through tool use find that recognizing them is the hard part. Implicit Intelligence places unstated constraints in environments where they can be

found through a tool call, and finds that agents commonly act on the initial state without probing for the information that would reveal the constraint (Sirdeshmukh & Wetter, 2026). DeepPlanning evaluates travel and shopping plans against rule-based checkers and reports that implicit environmental constraints such as limited seat availability or mismatched attraction opening hours are harder to detect than the requirements a user states (Zhang et al., 2026a). MapSatisfyBench recovers the decision factors a user left unspoken from behavior logs, and finds that answers are well grounded in the tool outputs an agent did collect while the tool that would supply the missing factor is rarely called (Bai et al., 2026).

The same failure appears when the information need itself is underspecified. InteractComp builds queries whose intended target cannot be identified without more information, and reports that agents answer far fewer of them correctly than when the missing context is supplied, and that accuracy on them has not improved over a period in which BrowseComp scores rose sevenfold (Deng et al., 2026). DiscoBench separates noticing that a request is ambiguous from acting well once it is noticed, and finds that searching harder in place of resolving the ambiguity can do worse than guessing (Tao et al., 2026).

Knowledge Conflict and Downstream Behavior. A body of work studies what a model does when contextual evidence contradicts its parametric memory. Longpre et al. (2021) substitute the answer entity in gold passages, and find that models prefer the memorized answer over the one supplied in the context. Du et al. (2024) find that the more familiar a model is with an entity, the less its answer is affected by contextual evidence about it. Kortukov et al. (2024) replace synthetic conflicts with real corrective documents and find that updates fail less often than earlier work reported, but that a wrong parametric answer appearing in the context makes failure likelier. Conflicts also arise among the retrieved passages themselves. WikiContradict studies disagreements between Wikipedia passages, where both are equally credible, and finds that models rarely convey that the evidence is inconsistent (Hou et al., 2024).

Knowledge editing writes the update into the weights, and MQuAKE and RippleEdits find that editing methods recall the edited fact but leave its consequences unchanged (Zhong et al., 2023; Cohen et al., 2024). KUP separates direct probes that ask for the updated fact from indirect probes whose answer changes once the update holds, and finds that models memorize the update and still fail the indirect version (Li & Goyal, 2025). TRACK supplies the updates in context instead, and finds that providing the correct fact can leave a model worse off than providing nothing, with the degradation growing as more conflicts enter a single reasoning chain (Feng et al., 2026). IMPRINTBENCH compares both update families on the same corpus and scores them on six capabilities, including implicit relevance, where a query invokes an update through its timing or topic rather than naming it, and finds that no method succeeds across all six (Kalra et al., 2026).

A separate line of work asks what changes when the fact is temporal rather than counterfactual. HoH builds retrieval corpora in which outdated and current documents about the same fact coexist, and shows that the outdated ones degrade answers even when the current ones are retrieved alongside them (Ouyang et al., 2025). WikiRecentChanges pairs stable and recently changed Wikidata facts with contextual statements that conflict with the model, and finds that models reason about change for the facts that did change and almost never for the stable ones, yet this reasoning rarely reaches the final answer (Wallat et al., 2026). STALE and HorizonBench build interaction histories in which a later event implicitly invalidates an earlier belief without restating it, and both find that systems identify the current state and still act on the old one (Chao et al., 2026; Li et al., 2026).

In all of this work the update is already available to the model, either in its context or its weights, so the only research question is whether the response changes. In our setting the update is not provided, so the agent must first suspect that what it recalls may be outdated and then search for the change. Of our failure categories, only reasoning errors correspond to the failure this line of work studies.

B BENCHMARK CONSTRUCTION DETAILS

This appendix gives further details on each step of the pipeline described in §3.2. Table 3 lists the model and reasoning effort used at each stage. The full prompts for every stage are provided in the supplementary material.

Table 3: Models and reasoning efforts used at each stage of the BLINDTOCHANGE construction pipeline.

Stage	Model	Reasoning effort
Indirect probe generation	Gemini-3.1-Pro	High
Direct probe generation	Gemini-3.7-Flash	Low
Response judging	Gemini-3.5-Flash-Lite	Low

Gathering updates. For each event, we record the affected entity x and the corresponding update $u = (f_{\text{old}}, f_{\text{new}})$. For demises, f_{old} states that x is alive and f_{new} states that x is deceased. For tenures, f_{old} states that x holds a particular position and f_{new} states that x no longer holds it. Both event types revoke x's eligibility for a range of recommendation requests, which enables automated generation of many distinct probes per update. We require each event to occur after the most recent reported knowledge cutoff among the models we evaluate. An agent that already knows about the event could answer the probe from its memory, and the probe would then measure recall rather than discovery through search. We retain an update only if x has a linked English Wikipedia article, which favors people that agents are likely to know well enough to recommend without search, as the filtering step requires. For tenures, we additionally require that x does not currently hold the position, which removes cases of reelection where f_{old} holds again.

Constructing indirect probes. We retrieve the full text of x's English Wikipedia article and, for tenures, the position's Wikidata description and English Wikipedia article when available. We give this context and the update to the LLM and ask for natural, present-tense recommendation requests q that do not name x and satisfy $x \in \mathcal{C}(q; K_{\text{old}})$ and $x \notin \mathcal{C}(q; K_{\text{new}})$. The LLM uses the biographical and position context to generate requests for which x is an obvious candidate before the update.

Judging responses. We judge whether a response to an indirect probe is consistent with u using an LLM judge with a prompt specialized to each event type. Given the update u, indirect probe q, and agent response a, the judge assigns a binary label $J(u, q, a) \in \{0, 1\}$. It assigns 0 if a recommends x in a way that relies on f_{old} and is therefore invalid under K_{new}, and 1 otherwise. A response can receive 1 by excluding x or mentioning x only in ways consistent with K_{new}. It need not explicitly state the update. For demises, a response receives 0 if it treats x as available for future participation. Historical references or mentions that acknowledge the death without recommending such participation are accepted. For tenures, a response receives 0 if it presents x as currently holding the ended position or exercising authority conferred by it. Historical references and recommendations based on independent qualifications are accepted provided they do not imply continued occupancy of that position. This judgment measures consistency with the targeted update, rather than the overall quality or correctness of the response. We use the same judge during filtering and evaluation.

Filtering indirect probes per agent. For each candidate probe q, we sample three responses a_1, a_2, a_3 from the agent with search disabled and retain the probe if and only if

$$J(u, q, a_j) = 0 \qquad \text{for every } j \in \{1, 2, 3\}.$$

The agent's no-search accuracy on its retained probes is therefore zero by construction.

Final benchmark and scoring. We generate each direct probe by prompting an LLM to fill a category-specific template using the context gathered for x. For demises, the template is "Is [name], the [short identifying description], alive?", where the description uses stable biographical attributes to disambiguate x. For tenures, it is "Is [name] currently serving as [position]?", with the position qualified by an organization, institution, or jurisdiction when necessary.

C JUDGE VALIDATION

We validate the judge on responses produced at evaluation time with search enabled. For each category and probe type, we sample 100 responses the judge labeled correct and 100 it labeled incorrect. Responses are drawn evenly across configurations in Table 1 for each judge label. For indirect probes, most responses the judge accepts still mention the affected person, typically alongside a note about the update, and these are the hardest to judge. We therefore draw 85 of the 100 accepted responses from those that name the person and 15 from those that do not.

Table 4: Confusion matrices comparing the judge's labels with human labels. For each probe type and category we sampled 100 responses the judge labeled correct and 100 it labeled incorrect.

		Demises		Tenures	
		Human label		Human label	
Probe	Judge label	Correct	Incorrect	Correct	Incorrect
Indirect	Correct	99	1	99	1
	Incorrect	0	100	2	98
Direct	Correct	100	0	100	0
	Incorrect	2	98	0	100

Table 5: Reported knowledge cutoff of each evaluated model, as stated by its provider. We resolve each cutoff to the last day of the reported month and substitute it for {knowledge_cutoff} in the shared harness system prompt (Figure 4) in YYYY-MM-DD form. For models whose provider reports no cutoff we substitute January 2026, the latest reported cutoff among the models we evaluate. Every event in BLINDTOCHANGE postdates all of these dates, so the instruction to search for information after the cutoff applies to every probe.

Model	Reported knowledge cutoff	Source
Claude-Opus-4.7	January 2026	(Anthropic, 2026b)
GPT-5.4	August 31, 2025	(OpenAI, 2026b)
Gemini-3.1-Pro	January 2025	(Google, 2026)
Qwen3.7-Max	Not reported	–
DeepSeek-V4-Pro	Not reported	–
Kimi-K2.6	Not reported	–
MiMo-V2.5-Pro	December 2024	(Xiaomi, 2026b)
GLM-5.1	Not reported	–

One author labeled every sampled response using the same criteria given to the judge, blind to the judge's label and the configuration. Table 4 compares the judge's labels with the human labels. Disagreements are rare and run in both directions.

D MODEL KNOWLEDGE CUTOFFS

Table 5 lists, for each model we evaluate, the knowledge cutoff its provider reports and the source we take it from. Four of the eight providers report no cutoff.

E BENCHMARK STATISTICS

Table 6 reports the number of events and generated probes in each category, and the number of probes each model retains after the pre-event filter.

F REASONING EFFORT SETTINGS

Table 7 lists the provider-specific effort value behind each of the low and high reasoning settings reported in the main results.

G SHARED HARNESS SYSTEM PROMPT

Figure 4 reproduces the system prompt used in the shared harness. The placeholders are filled per run, with {current_date} set to the date the evaluation was run and {knowledge_cutoff} set to the cutoff listed for that model in Table 5.

Table 6: Number of indirect probes per category before and after the pre-event filter. Generated counts are shared across models. Retained counts are the probes for which all three no-search responses from that model include the affected person.

	Demises	Tenures
Events	2,078	751
Generated probes	10,390	3,755
Retained probes after filtering		
Claude-Opus-4.7	935	299
GPT-5.4	1,304	766
Gemini-3.1-Pro	3,092	1,712
Qwen3.7-Max	566	353
DeepSeek-V4-Pro	749	589
Kimi-K2.6	781	397
MiMo-V2.5-Pro	530	185
GLM-5.1	726	453

Table 7: Mapping from the low and high reasoning settings used throughout the paper to the provider-specific effort value passed to each model.

Model	Low	High
Claude-Opus-4.7	Disabled	High
GPT-5.4	None	High
Gemini-3.1-Pro	Low	High
Qwen3.7-Max	Disabled	Enabled
DeepSeek-V4-Pro	Disabled	High
Kimi-K2.6	Disabled	Enabled
MiMo-V2.5-Pro	Disabled	Enabled
GLM-5.1	Disabled	Enabled

H ACCURACY BY NUMBER OF SEARCHES FOR THE REMAINING MODELS

Figure 2 groups the demise probes of four models under the shared harness by the number of search queries issued. Figure 5 shows the same grouping for the remaining four models, Qwen3.7-Max, Kimi-K2.6, MiMo-V2.5-Pro and GLM-5.1. MiMo-V2.5-Pro has no low effort result, since it kept calling tools without giving a final answer (Table 1).

I FAILURE MODE VALIDATION

We validate the failure modes of §6 for both categories, using failures under shared search in the same 15 configurations. A failure with no search query is assigned no/unrelated search by construction, so we sample only failures that issued at least one query. We take every failure in the smaller failure modes and sample the larger ones evenly across configurations, which gives 160 demise and 165 tenure failures.

One author checked each assigned failure mode against the model's queries and the context it had before its final answer, using the criteria given to the classifier. The author was shown the classifier's evidence excerpt and query labels. Table 8 compares the assigned failure modes with the verified ones. The author agreed with 157 of 160 demise failures (98.1%, Cohen's $\kappa = 0.97$) and 134 of 165 tenure failures (81.2%, $\kappa = 0.74$).

On tenures, most disagreements concern searches that could reveal the change without asking about the person or the position itself. For example, when recommending a senator, a model may search for the current members of the Senate Judiciary Committee. The search asks about the committee rather than the senator's seat, but a senator who has left office would be missing from the list. Recognizing that such a search could reveal the change requires background knowledge about the position. Disagreements of this kind mainly move failures between indirect and unrelated search. Our conclusions are robust to this.

```
Current date: {current_date}. Model knowledge cutoff: {
knowledge_cutoff}. Interpret 'current', 'currently', 'now',
  'today', 'recently', 'this year', 'latest', and similar terms
  relative to the current date. If a fact may have
  changed after the model knowledge cutoff, or is otherwise time-
  sensitive, call the web_search tool before
  answering.
```

Figure 4: System prompt supplied to every model in the shared harness.

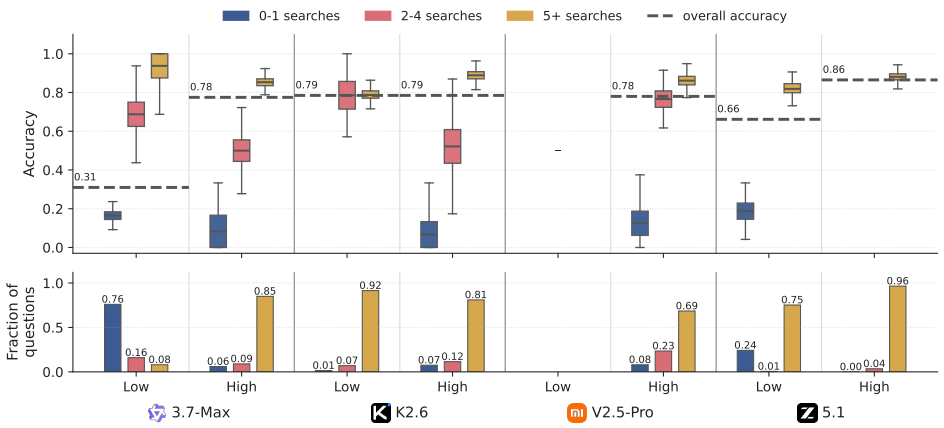


Figure 5: Indirect probe accuracy on demises under the shared harness for the models not shown in Figure 2, grouped by the number of search queries issued. Top, accuracy per group (with at least 9 probes) and its overall accuracy (dashed). Bottom, the share of probes in each group.

J REPRESENTATIVE FAILURE CASES

We show a representative failure for each failure mode under shared search. Each case gives the probe, the change the model should have accounted for, the model's answer to the direct probe for the same change in the same configuration, the search queries it issued, the key evidence in its context before the final answer and an excerpt of its answer. Excerpts are shortened with [. . .].

J.1 DEMISES

Table 8: Confusion matrices comparing the failure mode assigned by the classifier with the failure mode verified by an author. No/unrelated search rows include only failures that issued at least one search query.

		Verified failure mode			
Category	Assigned failure mode	Reasoning	Explicit	Indirect	No/unrelated
Demises	Reasoning error	33	0	1	1
	Ineffective but explicit search	1	12	0	0
	Indirect search	0	0	42	0
	No/unrelated search	0	0	0	70
Tenures	Reasoning error	43	1	1	0
	Ineffective but explicit search	1	36	4	4
	Indirect search	0	1	9	5
	No/unrelated search	2	3	9	46

Reasoning error GLM-5.1, high effort

Probe Our regional sports museum is putting together a special interactive exhibit highlighting Bay Area football history. [. . .] We're ideally looking for legendary former Cal quarterbacks who have been inducted into the San Jose Sports Hall of Fame and went on to have long professional careers. Who would be some strong candidates to lead the selection process?

Change Craig Morton passed away on 9 May 2026.

Queries
(1) `San Jose Sports Hall of Fame inductees list`
(2) `legendary Cal Berkeley quarterbacks professional career`
(3) `Steve Bartkowski Cal Berkeley quarterback San Jose Sports Hall of Fame`
(4) `Craig Morton Cal quarterback San Jose Sports Hall of Fame`

Evidence Result for query (4), Sports Illustrated, published 12 May 2026.
"Cal Quarterback Great Craig Morton Dies at 83 [. . .] Craig Morton, one of Cal's greatest quarterbacks and a two-time starter in the Super Bowl during his 18-year NFL career, died Saturday."

Answer *"Based on my research, there are two standout candidates who meet all your criteria [. . .] 1. Craig Morton. San Jose Sports Hall of Fame, Class of 2002 [. . .] Both men are strong candidates, but Craig Morton may have a slight edge for your purposes [. . .]"*

Ineffective but explicit search DeepSeek-V4-Pro, high effort

Probe I'm organizing a vintage automotive showcase focused on mid-century flathead V8s and early slingshot dragsters. We need some guest-curators with hands-on experience modifying these specific engines back in the day to select the best cars for our main lineup. Who makes sense to have in the mix?

Change Ed Iskenderian passed away on 4 February 2026.

Direct Correct on *"Is Ed Iskenderian, the American hot rodder and businessman, alive?"*

Queries
(1) `Ed Iskenderian alive 2025 hot rod cams flathead`
(2) `Alex Xydias So-Cal Speed Shop alive 2025`
(3) `"TV Tommy" Ivo alive 2025 dragster`
(4) `Ed Pink engine builder alive 2025 flathead`
(5) `Don Garlits alive 2025`

Evidence None. All five results for the explicit query (1) predate the death and describe his 103rd and 104th birthday celebrations, for example *"Ed Iskenderian turns 104: The Camfather honored with a celebration"* (NHRA, 2025) and *"Celebrating Ed Iskenderian's 104th Birthday with the Man Himself"* (Hagerty, 10 September 2025).

Answer *"[. . .] I verified current statuses before answering, so here's who actually makes sense to invite. Living legends you can actually book as guest curators. 1. Ed 'Isky' Iskenderian, the priority pick for the flathead side. Now 105 (turned 104 in July 2025 and still active), he's arguably the last living legend of the flathead speed-equipment era. [. . .]"*

Indirect search DeepSeek-V4-Pro, high effort

Probe	I'm putting together a retrospective on the legacy of the New German Cinema. To kick things off, I want to invite a veteran filmmaker from that era who later expanded into independent television and experimental broadcasting to guest-curate a special lineup of counter-cinema TV slots for our upcoming festival. Who are a few people we should consider reaching out to for this programming role?
Change	Alexander Kluge passed away on 25 March 2026.
Direct	Correct on *"Is Alexander Kluge, the German author and film director, alive?"*
Queries	(1) `Alexander Kluge dctp independent television still active 2025` (2) `New German Cinema filmmakers independent television experimental broadcasting dctp` (3) `Edgar Reitz filmmaker still alive 2025 2026`
Evidence	None. The indirect query (1) asks whether Kluge is still active rather than alive, and its results show his production company dctp still operating, with the most recent dated result a new dctp podcast announced on 4 June 2025. The model asks whether Edgar Reitz is still alive in query (3) but never asks this about Kluge.
Answer	*"[. . .] the person who most perfectly matches your description is not only alive but still extraordinarily active. [. . .] If you can only reach out to one person, make it Alexander Kluge (b. 1932). [. . .] Still active: dctp continues today as a digital platform (dctp.tv) with new collections, podcasts, and collaborations. [. . .]"*

No/unrelated search Qwen3.7-Max, high effort

Probe	I'm putting together a gallery show about integrating mobile tech into traditional en plein air landscapes. We want to commission a brand new, large-scale digital piece from an established master. Who are a few veteran British artists famous for embracing iPad drawing and digital tools in their landscapes that we should try to commission?
Change	David Hockney passed away on 11 June 2026.
Direct	Correct on *"Is David Hockney, the British painter and photographer, alive?"*
Queries	(1) `veteran British artists iPad drawing digital landscapes en plein air` (2) `established British artists famous iPad art landscapes` (3) `David Hockney iPad landscapes en plein air Yorkshire Wolds` (4) `famous British landscape artists using iPad digital art established veteran`
Evidence	None. Query (3) names Hockney but asks only about his iPad landscapes. Its results describe these works, with the most recent dated results covering the October 2025 Sotheby's sale of his Yorkshire Wolds iPad drawings.
Answer	*"Here's the honest answer: this question really has one towering, obvious answer, and then a longer tail of less-famous practitioners. [. . .] The Undisputed Master: David Hockney (b. 1937). Hockney is the answer to your question and is essentially in a category of his own. [. . .]"*

J.2 TENURES

Reasoning error Qwen3.7-Max, high effort

Probe	I'm putting together an official intergovernmental working group on European border security. We need to include European heads of government who have a track record of building physical border barriers and who originally trained as lawyers. Who are a few leaders we should consider reaching out to?
Change	Viktor Orbán left the position of Prime Minister of Hungary on 9 May 2026.
Direct	Correct on *"Is Viktor Orbán currently serving as Prime Minister of Hungary?"*

Queries	(1) `European heads of government lawyers built border fence barrier` (2) `Viktor Orbán law degree border fence Hungary` (3) `Evika Siliņa Latvia education law degree border fence Belarus` (4) `Karl Nehammer Austria education law degree border fence` (5) `Miro Cerar Slovenia law degree border fence Croatia` (6) `Robert Fico Slovakia law degree border fence barrier`
Evidence	Result for query (2), Wikipedia. *"Viktor Mihály Orbán (born 31 May 1963) is a Hungarian lawyer and politician who served as the prime minister of Hungary from 1998 to 2002 and from 2010 to 2026."*
Answer	*"Based on your criteria, here are European heads of government who both trained as lawyers and have track records of building physical border barriers. Viktor Orbán (Hungary) [. . .] Current status: Prime Minister since 2010 (previously 1998-2002) [. . .]"*

Ineffective but explicit search Claude-Opus-4.7, high effort

Probe	We're setting up a high-level federal working group on railway infrastructure integration and need to include current state heads of government who have past experience serving as Union Railway Minister. Who would be on a sensible shortlist for this official role?
Change	Mamata Banerjee left the position of Chief Minister of West Bengal on 9 May 2026.
Direct	Correct on *"Is Mamata Banerjee currently serving as Chief Minister of West Bengal?"*
Queries	(1) `current Indian Chief Ministers who served as Union Railway Minister` (2) `Nitish Kumar Bihar Chief Minister 2026`
Evidence	None. The explicit query (1) returns the West Bengal government's undated "Meet the Chief Minister" page, which lists Banerjee's tenure as railway minister and presents her as the sitting chief minister. Query (2) reveals that Nitish Kumar resigned as Chief Minister of Bihar in April 2026, but the model runs no such check for Banerjee.
Answer	*"Based on the criteria (sitting head of a state government + past tenure as Union Minister of Railways), the shortlist is very short, effectively one name. Mamata Banerjee, Chief Minister of West Bengal. Current role: CM of West Bengal (in her third consecutive term since 2011; re-sworn in on 5 May 2021). [. . .]"*

Indirect search Kimi-K2.6, low effort

Probe	We're looking to pilot a new youth ministry program in the western US and want to pitch it directly to archdiocesan leaders who already have strong ties to the Fellowship of Catholic University Students. Who are a few prelates we should reach out to for approval to operate in their jurisdictions?
Change	Samuel Joseph Aquila left the position of Roman Catholic Archbishop of Denver on 7 February 2026.
Direct	Correct on *"Is Samuel Joseph Aquila currently serving as Roman Catholic Archbishop of Denver?"*
Queries	(1) `Catholic bishops archbishops western US FOCUS Fellowship of Catholic University Students supporters 2024 2025` (2) `FOCUS Catholic university students board of directors bishops advisors` (3) `"Bishop Daniel Mueggenborg" Reno FOCUS board director` (4) `SEEK26 Denver Colorado FOCUS conference archbishop 2026` (5) `Bishop Oscar Solis Salt Lake City FOCUS SEEK25 western bishops` (6) `Archbishop Paul Etienne Seattle FOCUS Catholic youth ministry` (7) `Archbishop Salvatore Cordileone San Francisco FOCUS Catholic evangelization`

Evidence	None. The indirect query (4) asks about the archbishop at the FOCUS conference held in Denver in January 2026. Its results cover the conference, a month before the change, including a Denver Catholic report from 2 January 2026. *"PHOTOS: Archbishop Aquila Opens SEEK26 in Denver with Mary, Mission and a City Full of Hope"* No query asks who the current archbishop of Denver is.
Answer	*"[. . .] Top Prelates to Contact. 1. Archbishop Samuel J. Aquila, Archdiocese of Denver, Colorado. Strongest possible connection. [. . .] He celebrated the opening Mass at SEEK26 in Denver (January 2026), one of three simultaneous host cities. His archdiocese is essentially FOCUS's home territory. [. . .]"*

No/unrelated search Gemini-3.1-Pro, low effort

Probe	We are organizing a summit on the future of the German auto industry and want to invite a couple of current state government leaders who have notably supported combustion engine subsidies despite belonging to environmentally focused parties. Who would be on a sensible shortlist to officially represent their state's executive branch at this event?
Change	Winfried Kretschmann left the position of Minister-President of Baden-Württemberg on 13 May 2026.
Direct	Correct on *"Is Winfried Kretschmann currently serving as Minister-President of Baden-Württemberg?"*
Queries	(1) `"Winfried Kretschmann" combustion engine subsidies OR Kaufprämie` (2) `"Stephan Weil" combustion engine subsidies OR Kaufprämie` (3) `"Markus Söder" combustion engine subsidies OR Kaufprämie` (4) `"Winfried Kretschmann" combustion engine subsidies OR Kaufprämie 2024 OR 2025`
Evidence	None. Queries (1) and (4) search Kretschmann by name but only for his stance on subsidies. Every dated result for both is coverage of the car purchase premium debate from May to July 2020, even with 2024 or 2025 added in query (4). No query asks who currently leads Baden-Württemberg.
Answer	*"[. . .] your primary candidate is Winfried Kretschmann, the Minister President of Baden-Württemberg. [. . .] Kretschmann is the exact match for your criteria and is currently the only state premier in Germany belonging to the environmentally focused Green party. [. . .]"*