STOP OR CONTINUE: HOW FINE-TUNING CHANGES EVIDENCE USE IN SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

After every search, an agent must decide whether it has enough to answer or should keep searching. We ask which features of the retrieved page drive this decision, using matched edits that switch a mention of the question's entity, an answer-like string and the supporting fact on or off. Instruction-tuned models from three families are more likely to answer with an answer-like string when one appears, even if nothing supports it. After supervised fine-tuning on single answer-or-search decisions, Qwen2.5-7B-Instruct responds mainly to the mention instead: the effect of a mention on its rate of stopping (answering or declining) rises from 6 to 47 percentage points and that of an answer-like string falls from 29 to 5, a reversal no shift in its overall willingness to stop can produce. It held on preregistered held-out pages; Mistral-7B-Instruct-v0.3, fine-tuned the same way, showed it on development pages. Asked about pages that mention the entity but lack the fact, the fine-tuned agent says the fact is missing for 96% of them but answers on 54%. In real search episodes, fine-tuned agents often answer once the page about the question's entity arrives, even on questions this page cannot answer, where every answer was wrong. Instructions to check for the fact barely changed what the fine-tuned agent stops on. Unlike public search-agent trajectories, its training data almost never contain a page about the entity without the fact, a coupling measurable before training; in a preregistered study on new questions, counter-examples of this kind removed 82% of the mention's effect on stopping, while generic negatives had no measurable effect. Four larger open models rarely stopped on either cue alone, but with reasoning off, five of six models from five families often answered with the answer-like string on pages linking the right entities by the wrong relation.

1 INTRODUCTION

Suppose a search agent is asked where the first woman to win a Nobel Prize was born. Its first search finds Marie Curie. Its second returns a page about her that does not say where she was born. The agent now has to decide whether to answer or to search again, and it faces this decision after every result it reads. Stopping too early gives an answer that the evidence does not support, and stopping too late wastes calls. Deployed search tools often make the first mistake. They rarely decline questions they cannot answer accurately (Jaźwińska & Chandrasekar, 2025), and many statements in their answers are not supported by the sources they cite (Liu et al., 2023; Venkit et al., 2026). In multi-hop search the page that names the right entity is often not the page that states the needed fact, and an agent that answers from the first page stops one hop too early.

Ideally the decision would depend on support, that is, on whether the retrieved text states the fact the question asks for (sufficient context in the terms of Joren et al., 2025). But pages that state the fact usually also name the entity the question asks about and contain a string of the answer's type, and pages that differ in these features also differ in many other ways. Natural data therefore cannot tell which feature an agent responds to. We separate the features by editing single retrieved pages, as in contrast sets (Gardner et al., 2020). The mention of the entity and the answer-like string are each switched on or off, with and without the fact, while the rest of the page stays the same (Figure 1). We then look for the same behavior in real search episodes.

Features other than support move the decision substantially, and which of them matters most changes with fine-tuning. In instruction-tuned models from three families (Qwen, Mistral and Gemma),

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.

Question: Where was Marie Curie born?

neither	*Pierre* Curie shared the 1903 Nobel Prize in Physics.
mention	*Marie* Curie shared the 1903 Nobel Prize in Physics.
candidate	Pierre Curie shared the 1903 Nobel Prize in Physics, presented in *Stockholm*.
fact	Marie Curie, *born in Warsaw*, shared the 1903 Nobel Prize in Physics.

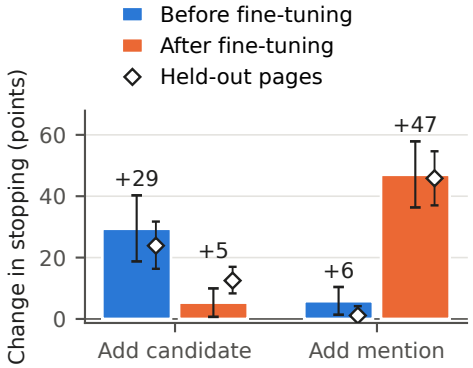

Figure 1: What makes a search agent stop. *Left:* an illustrative retrieved page in four versions; the *fact* version necessarily contains the mention and a candidate. The passages in our experiments are longer, and the questions ask about invented entities (Appendix A). *Right:* change in the rate of stopping (answering or declining), in percentage points, when a mention of the subject or an answer-like candidate is added to pages without the fact, averaged over the presence of the other feature, for Qwen2.5-7B-Instruct before and after our fine-tuning. Before fine-tuning the model reacts mainly to the candidate, after it mainly to the mention. Bars show the development pages and diamonds the held-out pages, which were evaluated once under a preregistered plan; error bars are 95% bootstrap intervals over question clusters.

adding an answer-like string that nothing on the page supports raises the rate of answering with it by 10 to 58 percentage points across six model–dataset combinations (Section 3). For Qwen the same holds with code-search and SQL tools, and in unedited runs most of its answers that do not match the gold answer are strings from the retrieved text.

After supervised fine-tuning on single answer-or-search decisions, Qwen2.5-7B-Instruct responds mainly to a different feature. On our edited pages, mentioning the question's entity raises its stopping rate by 47 percentage points and adding an answer-like string by 5, compared with 6 and 29 before fine-tuning (Figure 1). On the log-odds of stopping, the mention effect grew while the candidate effect shrank, which a uniform shift or rescaling of the willingness to stop cannot produce. All five runs in Table 2 trained without counter-examples show this pattern, a preregistered test on held-out pages confirmed it, and the same recipe reverses the cues of Mistral-7B-Instruct-v0.3 as well. In a plain prompt, the agent judges the fact missing on 96% of the pages that mention the entity without it, yet it answers on 54% of them (Section 4.2). In real search episodes, fine-tuned agents often answer as soon as they have retrieved the page about the entity the question names, before the page that states the needed fact. On questions that this page cannot answer, they stop about as often, and none of these answers is correct (Section 4.3).

A likely reason lies in the training data. In the examples we fine-tune on, whether the retrieved text includes a page about the question's entity predicts the target action almost perfectly (area under the ROC curve, AUROC, 0.98), and no example asks for a search once the first hop has been found, so in these data a mention almost always comes with the fact, and stopping on the fact is hard to tell from stopping on a page about the entity. Public training trajectories of search agents contain the missing states in about half of their search examples and show little coupling (AUROC 0.52 to 0.56; Section 4.4), so the statistic can flag this risk before training. When two features predict the label, models tend to learn the one that is easier to extract (Geirhos et al., 2020; Hermann et al., 2024), and a name is likely easier to spot than a relation is to check. Telling the agent to answer only when the text states the needed fact made it more cautious overall but barely changed what it stops on. By contrast, in a preregistered study on edited pages built from new questions, counter-examples in the training data that mention the entity without the fact removed 82% of the dependence of stopping on the mention, while generic negative examples had no measurable effect (Section 4.5).

Recent work finds that deep search agents both search redundantly and stop too early, and attributes both to misaligned decision boundaries (Zhang et al., 2026a); concurrent work shows that search agents adopt answers from evidence written for a nearby question (Zhang et al., 2026b). We contribute

a controlled test of which page features the stop decision responds to; evidence that a fine-tuned agent stops on a mention even when it can tell that the page lacks the fact; a statistic of the training data, computable before training, that tracked the learned feature across our runs; and a change to the training data that reduced the mention's effect in a preregistered test on new questions. The failures we describe are not universal: four larger open models we tested hardly ever stopped on a page with only a mention or only an unsupported answer-like string. With reasoning turned off, however, five of six models from five families answered with the candidate on at least a third of one-hop pages that link the right entities by the wrong relation, and none responded more to the mention than to the candidate (Section 4.6).

2 SETUP

We study one decision: whether, right after a search result, the agent answers, searches again or declines. We call the text that one search returns a *page*; here a page holds one or more short passages.

Three features of a retrieved page. Consider the question "Where was Marie Curie born?" We call Marie Curie the question's *subject*, the entity whose property is asked. In a multi-hop question the subject of a later hop is a *bridge entity* found at an earlier hop, as in the example of Section 1. A retrieved page can *mention the subject* ("Marie Curie shared the 1903 Nobel Prize in Physics"). It can contain an *answer-like candidate*, a string of the answer's type such as the name of a city ("... presented in Stockholm"). And it can state the *supporting fact*, the sentence that links the subject to the answer through the asked relation ("Marie Curie was born in Warsaw"). The fact usually comes with the other two features, so natural pages confound them. Pages with only some of the features are common in multi-hop search: after the first search, the page about the entity the question names is often retrieved without all passages that state the needed facts (199 of 305 questions in Appendix D), and pages about related entities often contain strings of the answer's type (Section 3).

Outcomes. We record the agent's next action. A tool call is a *search*. A final message is an *answer* if it commits to an answer and a *decline* if it says that the information is insufficient. An LLM judge (Zheng et al., 2023), blind to the condition, labels every final message, and a second judge relabels them to measure agreement (Cohen's $\kappa = 0.90$; Appendix A). The agent *stops* when it sends a final message, almost always an answer or a decline; whether it stops is read from the output format and does not depend on the judges. For pages with a candidate, we also record whether the answer is the candidate. Besides actions, we record the *log-odds of stopping*: the difference between the log-probabilities of starting a final answer and starting a tool call at the decision point.

Candidate pages. To compare instruction-tuned models across families, we use a design in which only the candidate is manipulated. We build matched triples from MuSiQue (Trivedi et al., 2022), a multi-hop benchmark built from Wikipedia that remains in use for training and evaluating search agents (Chen et al., 2025; Jin et al., 2025), and from 50 items generated from eight templates in the style of HotpotQA bridge questions (Yang et al., 2018), with invented entities. Following entity-substitution studies of context use (Longpre et al., 2021), the page returned by the agent's first search either has no candidate, contains a clause with a candidate that nothing on the page connects to the question, or also contains a sentence that supports that candidate. Comparing the first two pages measures the effect of adding the candidate clause. Each model sees 50 triples per set under one prompt, parser and scoring rule, with greedy decoding (Appendix C).

Three-switch pages. For the fine-tuning study we manipulate all three features. We start from two-hop questions of 2WikiMultiHopQA (Ho et al., 2020) that ask for an attribute of a person linked to the question's entity (its father, mother, spouse or director), and replace that entity with an invented name, so that without the fact a question has no correct answer. Each passage has a name slot that holds either the question's entity or another invented entity, and a candidate slot that holds the original question's answer or is empty, and it either states the fact or does not; the rest of the passage is real Wikipedia text. Each passage is shown with two questions, one about each invented entity, so the effect of a mention is measured on identical text. We also build *relation-swapped* pairs that keep the subject and the candidate and change only the relation word, as in "Marie Curie was born in Warsaw" versus "Marie Curie studied in Warsaw"; only the first answers the question, and we call the second a *near miss*. On all three-switch pages, models before and after fine-tuning receive

Table 1: Effect of an unsupported candidate before fine-tuning. Each cell is the change in the rate of answering with the candidate when an unsupported candidate is added to the page, paired over 50 questions, with the exact McNemar p-value in parentheses.

Model	Templates	MuSiQue
Qwen2.5-7B-Instruct	+0.32 (3.1×10^{-5})	+0.44 (4.8×10^{-7})
Mistral-7B-Instruct-v0.3	+0.58 (3.7×10^{-9})	+0.32 (4.0×10^{-4})
Gemma-2-9B-it	+0.10 (0.063)	+0.26 (2.4×10^{-4})

identical token sequences. Questions are screened by two LLM reviewers and pages are checked automatically for length and form; a sensitivity analysis excludes questions with possible real-world cues (Appendix A). The main analyses use 896 pages built from 56 parent questions of a development split; a held-out split of 1,328 pages from 83 further parent questions was evaluated once, for a preregistered confirmation.

Asking and instructing. To test whether an agent can tell that a page lacks the fact, we also ask it directly, in a plain prompt without tools, whether the passage states the fact the question needs, and compare its answer with its action on the same page. To test whether instructions change what an agent stops on, we append one of three instructions to its system prompt: one that describes a page about the entity without the fact, a reworded version, and a generic instruction to answer only when the text contains enough information (wording in Appendix B). The instructions and the rules for reading both analyses were fixed before the runs.

Real search episodes. Edited pages show what each feature does but not how often the situation arises. For Qwen2.5-7B-Instruct before fine-tuning we use 500 states right after its first search, in its own prompt format (Appendix C); for the fine-tuned agents, the first decision of real search episodes on 600 multi-hop questions, whose annotations tell us whether the first results include the page about the named entity and all supporting passages (Appendix D).

Models. The instruction-tuned models are Qwen2.5-7B-Instruct (Qwen Team, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and Gemma-2-9B-it (Gemma Team, 2024); we also ran Llama-3.1-8B-Instruct (Llama Team, 2024), reported in Appendix C. The fine-tuned agents start from Qwen2.5-7B-Instruct, except two from Mistral-7B-Instruct-v0.3, and are trained with supervised learning using low-rank adapters (Hu et al., 2022), as in trajectory fine-tuning of language agents (Chen et al., 2023) but on single decisions. Each training example pairs a multi-hop question from MuSiQue or 2WikiMultiHopQA with one retrieved observation. When the observation contains the passages that state the answer, the target is that answer; otherwise it is a search for the next bridge entity, whose name is taken from the dataset's annotations (706 examples from 239 questions; Appendix E). The main agent is trained on these examples. A second set of runs keeps the same examples and turns increasing shares of the search examples into counter-examples by replacing one of their passages; the versions with none and with all 144 possible counter-examples were trained with three seeds. Section 4.5 describes a separate training study, and Section 4.6 eight recent models from six families, accessed through APIs.

Statistics. Unless stated otherwise, intervals are 95% bootstrap intervals that resample clusters of related questions; paired binary comparisons use exact McNemar tests. For effects of fine-tuning the unit of replication is the training run; Table 2 reports the Qwen runs evaluated on the development pages, and Appendix E all runs, including those of the separate study in Section 4.5.

3 Before Fine-Tuning: Answer-Like Candidates

Adding an unsupported candidate to the page makes instruction-tuned models more likely to answer with it (Table 1). The effect is positive in all six model–dataset combinations. It is smallest for Gemma on the template items, the only combination where it is not significant. Without the candidate on the page, Qwen gives it as its answer in 0% of template and 2% of MuSiQue cases, so the increase is not a prior guess that the page merely confirms.

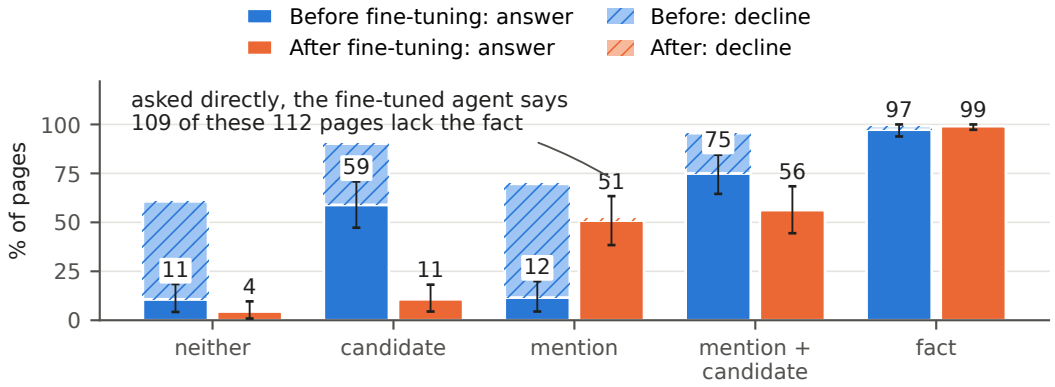

Figure 2: Answers and declines on three-switch pages. For each page type, the solid part of a bar shows how often the model answers and the hatched part how often it declines; the rest of the time it searches or gives a message that is neither. On pages that mention the subject but contain no candidate and no fact, the fine-tuned agent answers much more often than before fine-tuning and almost never declines, although, asked directly, it says that 109 of these 112 pages lack the fact (Section 4.2). Development split; 95% bootstrap intervals over question clusters.

The candidate effect also appears on 2WikiMultiHopQA for Qwen and Mistral (Gemma: +0.10, $p = 0.063$), under sampled decoding for Qwen, in three larger Qwen models on MuSiQue under an earlier version of the prompt, and for Qwen with two other tools. When a code-search result offers a legacy file whose name matches the query, or a prefilled SQL query computes the wrong metric but returns a row that looks like an answer, Qwen commits to that result 60 and 78 percentage points more often than when no such result is shown (Appendix C).

Answering with a string from the page also happens without planted candidates. Of the 293 answers that Qwen gave right after its first search in 500 unedited states and that neither match nor contain the gold answer, 219 are strings from the retrieved text, 121 of them while the gold answer is not in the text, the situation that the candidate pages create on purpose (Appendix C).

On three-switch pages, Qwen before fine-tuning shows the same preference. Adding a candidate raises its stopping rate by 29 percentage points, and mentioning the subject by 6 (Figure 1). When a page lacks the fact, it often declines instead of answering (Appendix B).

4 After Fine-Tuning: Subject Mentions

After fine-tuning, a mention of the subject matters much more for stopping than an answer-like candidate, the reverse of the pattern before fine-tuning.

4.1 Edited pages

Mentioning the subject raises the fine-tuned agent's stopping rate by 47 percentage points and adding a candidate raises it by 5; before fine-tuning the same edits gave 6 and 29 (Figure 1). A general change in the willingness to stop would leave both effects unchanged or rescale them by the same factor on the log-odds scale. Instead they moved in opposite directions, and the mention's share of the combined log-odds effect rose from 0.05 to 0.89 (Appendix B). All five runs in Table 2 trained without counter-examples show the pattern: in each, the mention raises stopping by more than the candidate does (34–52 against 5–17 percentage points), with an interval for the difference that excludes zero. Mistral-7B-Instruct-v0.3 fine-tuned once with the same data and recipe shows the same reversal: the mention raises its stopping rate by 35 percentage points and the candidate by 0, against 4 and 40 before fine-tuning, and with all counter-examples the mention effect is 22 (Appendix E). On the held-out pages, evaluated once after the analysis was fixed, adding the mention raised the fine-tuned agent's stopping rate by 46 [37, 55] percentage points and adding the candidate by 13 [8, 17], against 1 [−2, 4] and 24 [16, 32] before fine-tuning; both preregistered comparisons held (Appendix B).

Because the model often declines before fine-tuning, its stopping rate mixes two outcomes. Counting only answers gives the same picture (Figure 2). On pages that mention the subject but contain neither a candidate nor the fact, the fine-tuned agent answers in 51% of cases, against 12% before fine-tuning. Because the entity the question asks about is invented, only the missing fact could support an answer to these questions. The fine-tuned agent rarely declines (6 of 942 final messages); none of its training examples is a decline. On pages with the mention and the candidate but without the fact, nearly all its answers are the candidate (59 of 63), although adding the candidate barely changes how often it stops. When the page states the fact, the agent answers in 99% of cases, and in relation-swapped pairs, changing only the relation word lowers its stopping rate by 52 percentage points on one-hop and two-hop pairs alike. The agent can thus use the relation when everything else is held fixed, but a mention alone is often enough for it to stop.

4.2 Asking versus acting

Asked in a plain prompt, the fine-tuned agent can tell when a page that mentions the subject does not state the fact. On the 224 such pages, with or without the candidate, it says the passage lacks the needed fact for 96% of them, but as an agent it answers on 54% of the same pages. The mention's effect on stopping exceeds its effect on saying "yes" by 0.44 [0.34, 0.56], and by 0.21 [0.11, 0.31] when the two outcomes are matched for base rate. The same holds in the agent's own input format, where we force a choice between "Yes" and "No" (exploratory: the plan made this format primary but prescribed the plain prompt if fewer than 95% of its answers were valid, and only 78% were). The agent judges the fact missing on 87% of these pages, 92 of its 121 stops on them come on pages it judged that way, and the gap between the mention's effect on stopping and on the judgment is 0.37 after fine-tuning against 0.08 before (Appendix B).

4.3 Real search episodes

Real search episodes show the same early stop. We took the first decision of five fine-tuned runs on 600 multi-hop questions. The page about the entity a question names usually supplies only the first hop; the fact that answers the question is on another page. Across runs, 47–57% of first-step answers came before all supporting passages had been retrieved. When the first results included the page about the named entity but not all supporting passages, the runs stopped in 30–60% of cases, and 33–36% of these answers were correct. To check whether the agents stop in these states only when they can know the answer, we used 200 further questions, selected because one of the runs answers them wrongly when shown only the first-hop page. In the same states, the main agent and the two runs of the second set with seed 42, without and with counter-examples, stopped about as often as on the main questions (52%, 42% and 35%, against 54%, 42% and 30%), and none of their 61, 50 and 41 answers was correct; only the second of these runs was used to select the questions.

Across first decisions made before all supporting passages were retrieved, the page about the named entity raised the log-odds of stopping for the main agent (1.03 [0.44, 1.61]) and by less, not significantly, for the run with counter-examples (0.16 $[-0.42, 0.73]$); because that page also supplies the first hop, this exploratory comparison mixes the mention with partial evidence (Appendix D).

4.4 Why the agent learns the mention

Across our runs, the mention effect follows how well a mention predicts stopping in the training data (Table 2). In the examples used to train the main agent, whether the observation includes a page about the question's entity predicts the target action with an AUROC of 0.98, and whether it states the fact determines the target by construction. None of the 473 examples whose target is a search includes the page of the first hop, so the data never show a state in which the first hop has been found and the second is missing. The search targets are bridge entities taken from the dataset's annotations; for 234 of them the name appears nowhere in the observation (Appendix E). The data therefore almost never separate a mention from the fact. An agent that answers whenever the entity's page is present fits them almost as well as one that checks for the fact, and spotting a name is plausibly easier than checking a relation. Public data are far less coupled: the AUROC is 0.61 for MuSiQue with its unanswerable questions (Trivedi et al., 2022), 0.56 for the preference data of Zhang et al. (2026a) (by mentions, as its pages have no titles), and 0.52 to 0.56 for three training sets of search agents, in which 48–50% of the search examples come after the first-hop page and before the second (Appendix E). Our account

Table 2: Fine-tuning runs on the development pages. Coupling is the AUROC of "the observation includes a passage titled with the subject" as a predictor of the stop label in the training data (for the preference data, which have no titles, a mention of the subject). Effects are changes in the log-odds of stopping on three-switch pages. A control trained on permuted targets never stops and is not shown.

Training data	Seed	Coupling	Mention effect	Candidate effect
Main agent's training data	42	0.98	5.12	0.60
same data, earlier trainer	42	0.98	4.13	1.42
Second set,	42		4.74	1.21
no counter-examples	43	0.98	4.26	0.86
	44		3.78	1.68
Second set, 11% of counter-examples	42	0.96	4.32	1.44
Second set, 37% of counter-examples	42	0.90	3.16	1.54
Second set,	42		2.74	1.32
all counter-examples	43	0.76	2.28	1.13
	44		2.33	2.00
Public: preference data (Zhang et al., 2026a)	–	0.56	–	–
Public: MuSiQue, answerable and unanswerable	–	0.61	–	–
Public: three training sets of search agents	–	0.52 to 0.56	–	–

predicts a weaker mention cue in agents trained on these data. In R3-RAG's cold-start model, trained by its authors on one of these sets, the mention raised the stopping rate on our held-out pages by 14 percentage points and the candidate by 52 (Appendix F).

4.5 Instructions versus training data

Instructions barely changed what the fine-tuned agents stop on. After matching overall stopping rates, the first instruction changed the main agent's mention effect by only -0.044 [-0.093, $+0.002$] (raw change -0.141), and none of the three instructions reduced the mention effect of the runs without counter-examples; some made the agents more cautious overall, so the agents did not simply ignore them. Before fine-tuning, Qwen reacted more strongly, stopping on 42% of pages instead of 88%, but 98–99% of that change is reproduced by shifting its stop threshold alone. The effect of the fact stayed positive under every instruction (Appendix B). Prompting a search agent to check what the evidence supports reduces, but does not remove, its adoption of answers to nearby questions (Zhang et al., 2026b). Starting the main agent's reply with "Thought:" did lower its mention effect, from 47 to 11 points (not matched for overall stopping, which also fell), but not the candidate effect of Qwen before fine-tuning (exploratory; Appendix B).

Counter-examples in the training data did reduce the mention effect. In the second set of runs, one passage in some search examples is replaced by a passage that mentions the subject but lacks its fact sentence, while labels and the amount of data stay the same. Like counterfactually augmented data (Kaushik et al., 2020), these counter-examples break the link between the mention and the label. From the version without counter-examples to the version with all of them, the mention coupling falls from 0.98 to 0.76, and the mention effect falls from 4.26 to 2.45 logits (means over three seeds; Table 2). For seed 42, the two intermediate versions lie in between (4.32 and 3.16, against 4.74 and 2.74 at the ends). The candidate effect rose slightly, from 1.25 to 1.48, in all three seed pairs, possibly because deleting the fact sentence also deleted the answer string; removing one shortcut can strengthen another (Li et al., 2023). After shifting the stop threshold of the version without counter-examples to the overall stopping rate of the version with all of them, the mention effect on stopping was still lower with counter-examples, by 0.16 [0.10, 0.24] for seed 42 and 0.13 pooled over three seeds, and by 0.12 [0.07, 0.18] on the held-out pages, where the preregistered test held for all three seeds. The threshold shift accounts for only 42% of the pooled drop, and the effects of support and of the candidate changed by less than 10 percentage points (point estimates).

A separate study tested counter-examples on new questions under an analysis plan fixed in advance. Its arms differ only in one passage of the same 144 search examples: an LLM-written targeted counter-example that mentions the subject but does not state the fact, a generic negative on another topic rewritten by the same LLM to a similar length, or neither. Three seeds per arm were evaluated on pages built from 61 new questions. We measure the dependence of stopping on the mention as the difference in stopping rate between pages that mention the entity asked about and pages that mention the other entity, summed over the two questions of a pair (range -2 to 2); without counter-examples it is 0.74 [0.55, 0.93]. Targeted counter-examples changed it by -0.61 [-0.77, -0.45], removing 82% of it, and by -0.58 [-0.75, -0.41] relative to generic negatives, which themselves had no measurable effect (-0.03 [-0.11, 0.04]). With seeds instead of questions as the unit, these two differences remain significant, by a narrow margin. The deletion-based counter-examples above did about as well (difference from the targeted ones $+0.03$ [-0.04, 0.09]). Correctness and the rate of extra searches changed by less than the preregistered bound of 0.15 (-0.028 and $+0.040$), although an exploratory analysis found a larger drop in correctness, 0.078, on answerable questions about entities seen in training (Appendix E).

4.6 Recent models, with and without reasoning

We also ran the three-switch pages through four larger open-weight models, with their default thinking mode on: DeepSeek-V4-Pro (DeepSeek-AI, 2026), the model served as deepseek-flash, and GLM-5.3 and GLM-5.3-Flash (GLM-5 Team, 2026); details are in Appendix C. None of them stopped on pages that mention the subject but contain neither the candidate nor the fact (0 of 110–112 pages per model), and on pages with a mention and a candidate but no fact they stopped in 1–7%. Stating the asked relation raised their stopping by 47–94 percentage points on one-hop and by 68–90 on two-hop relation-swapped pairs. GLM-5.3-Flash answered with the candidate on 27% of near misses, the other three on 2–4%. Because these models rarely stop on a bare mention or an unsupported candidate, the pages cannot show whether weaker versions of the two cues exist in them.

Two of the models, DeepSeek-V4-Pro and deepseek-flash, also ran with thinking turned off; GLM-5.3 rejects this setting and GLM-5.3-Flash ignores it. Without thinking, both answered with the candidate on most one-hop relation-swapped pages (84 and 108 of 112, against 3 and 5 with thinking), and stating the asked relation raised their stopping rate by only 26 and 4 percentage points on one-hop pairs, against 94 and 81 with thinking. By a rule fixed before this run, which compares the mention and candidate effects (Appendix C), both now resemble Qwen before fine-tuning rather than the fine-tuned agent. Without thinking, both also stopped on some pages with only a mention (22 and 50 of 112), but by their wording almost all of these stops were declines, and the mention effect did not grow. Rerunning three of the page types at temperature 1.0 instead of 0 changed the preregistered rates by at most 3 points. With reasoning turned off, four models from other families answered with the candidate on 108 (Kimi K3), 108 (MiniMax M3), 40 (Mistral Medium 3.5) and 13 (Qwen3.8 Flash) of 112 one-hop near misses, and by the same rule none resembles the fine-tuned agent; with its default reasoning, MiniMax M3 did so on 47 (Appendix C).

5 Related Work

When search agents stop. Search agents interleave reasoning and retrieval (Yao et al., 2023) and are trained with reinforcement learning (Jin et al., 2025; Gao et al., 2026) or with supervised learning on trajectories (Chen et al., 2023; Deng et al., 2026). Zhang et al. (2026a) find both redundant searching and premature stopping in deep search agents and attribute them to misaligned decision boundaries. Luo (2026) trains a separate judge that decides when Search-R1 should stop. Deng et al. (2026) show that shortcuts in training tasks let agents answer with less search than intended, and concurrent work shows that search agents adopt answers from correct evidence written for a nearby question (Zhang et al., 2026b). Liu et al. (2025) teach agents when to stop gathering information with pairs of trajectories that differ minimally in whether stopping is appropriate; our counter-examples are pairs of this kind, and the coupling in Section 4.4 explains why they help.

Using retrieved context. Models often answer when the retrieved context does not contain enough information (Joren et al., 2025). Entity substitution exposes conflicts between retrieved and memorized knowledge (Longpre et al., 2021); models readily accept coherent external evidence even when

it contradicts what they know (Xie et al., 2024) and can be misled by irrelevant context (Jia & Liang, 2017; Shi et al., 2023; Yoran et al., 2024). Methods that interleave retrieval with reasoning or learn when to retrieve make the same decision inside the policy (Trivedi et al., 2023; Asai et al., 2024). Pages that name a bridge entity can be essential for continuing a multi-hop search even when they look useless to a reader (Mukhopadhyay et al., 2026); these are the pages on which our fine-tuned agents stop too early.

Shortcut learning. Models learn features that predict the label in their training data even when those features were never intended (McCoy et al., 2019; Geirhos et al., 2020), and which of several predictive features they learn depends on how predictive and how easy to extract each one is (Lovering et al., 2021; Hermann et al., 2024). In multi-hop question answering, models often answer without connecting the supporting facts (Min et al., 2019; Trivedi et al., 2020). Counterfactually augmented training data reduce reliance on such features (Kaushik et al., 2020), but removing one shortcut can amplify another (Li et al., 2023). Our fine-tuning results are an instance of shortcut learning.

Knowing versus acting. Models can often assess whether they know an answer (Kadavath et al., 2022) and encode information about the truth of their outputs that the outputs do not reflect (Orgad et al., 2025), and LLM agents can fail to act on knowledge they state (Schmied et al., 2026). Our asking-versus-acting result is a case of this gap in search.

6 Discussion and Limitations

Fine-tuning changed which feature the agent stops on, and the fine-tuned agent stops on a mention even when it can tell that the fact is missing. Both fit an agent that learned to answer whenever a page about the entity is present, because in its training data that page almost always came with the fact.

The coupling between a feature and the stop label can be measured before any model is trained, and when a mention predicts the label almost perfectly, as in our data, we would expect an agent to learn the mention. In our experiments, counter-examples that mention the entity without the fact reduced this cue, also on new questions, while instructions and generic negative examples did not. Building counter-examples by deleting the fact sentence also deletes the answer string, which may explain why answer-like strings became a slightly stronger cue in our runs; relation-swapped pages would avoid this, but we have not tested them as training data. Asking a model whether the evidence suffices can overstate how well it stops. Our agent judged these pages well and still answered on them, so evaluations of stopping should score what agents do on such pages.

The four larger open models we tested almost never stopped on a bare mention or an unsupported candidate, but near misses misled one of them and, with reasoning turned off, most of the models we could run that way (Section 4.6). Unlike our fine-tuned agent, none of them responded more to the mention than to the candidate, which fits our account of where the mention cue comes from. Reasoning reduced these errors in all three models we ran both ways and lowered our agent's mention effect, but not the candidate effect of Qwen before fine-tuning. Similar near misses mislead search agents (Zhang et al., 2026b), so agents that decide without reasoning should be tested on them.

Limitations. Our fine-tuning results come from one supervised recipe, applied to Qwen2.5-7B-Instruct and, with one seed, to Mistral-7B-Instruct-v0.3, in multi-hop question answering; the training examples are single decisions built from dataset annotations and lack states that are common in public trajectory data (Section 4.4), and we tested only one agent trained on such data. Apart from the main comparisons, which were confirmed on held-out pages, the edited-page analyses use a development split. We did not study reinforcement learning, and public agents trained with it differ from ours in format, data and reward; the coupling statistic of Section 4.4 is one way to form an expectation from their training data. We did not test the proprietary models of OpenAI, Anthropic or Google. The real episodes come from the first step of an earlier evaluation loop, and we measure decisions at single states rather than end-to-end accuracy. At that first step, the one run with counter-examples we tested stopped less often than its counterpart without them (44% versus 53%), but its answers were not more often correct (53% in both), so we have not shown that reducing the cue improves accuracy. Outside question answering we tested only models without our fine-tuning, on two small tool tasks; agent benchmarks with richer tools, such as τ^2-bench (Barres et al., 2026), remain untested.

AI USE STATEMENT

In this work, we used generative AI tools to help develop the conceptual framework and refine hypotheses, design and implement the experiments, run the experiments and statistical analyses, interpret results, generate synthetic data (the invented entity names, the edited pages and the counter-examples), clean and reformat datasets, assist with translation, and annotate model outputs: LLM reviewers screened questions and passages, and two LLM judges labeled final messages (Appendices A and C). Whether an agent stops is read from its output format and does not depend on these judges. The remaining tasks that require disclosure, formulating and proving mathematical claims, are not applicable to this work. Additionally, we used generative AI tools to search for literature and data, create figures, draft and edit the text, and format references. Reported numbers were checked against the experiment records, and references against their sources, with LLM assistance in a pass separate from the analysis. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects. The edited pages combine public Wikipedia text, including passages about real public figures, with invented entities, so they attribute invented works and relations to real people; we will mark them as synthetic test material when we release them. A few invented names coincide with names of unrelated real entities (Appendix A), and no statement about those entities is intended. The released pages reuse Wikipedia text from MuSiQue and 2WikiMultiHopQA; the other public datasets were used only to compute statistics and are not redistributed. The failures we describe could be used to write pages that mislead search agents; we report them so that such agents can be tested and trained against them.

REPRODUCIBILITY STATEMENT

Appendices A to E describe how every edited page, episode and fine-tuning set was built, with the instructions, the judging setup, the automatic checks and the training settings. We will release the code with the full prompts, both splits of the edited pages (the held-out split has now been used for its confirmation), the judge prompts and the fine-tuning configurations. The analysis plans of the preregistered comparisons were written before the corresponding runs and recorded with file hashes; we will release them with the hashes. Rerunning the development evaluation on a second machine gave identical outputs for all 7,168 records and a byte-identical results file.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating conversational agents in a dual-control environment. In *International Conference on Machine Learning (ICML)*, 2026.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. RQ-RAG: Learning to refine queries for retrieval augmented generation. In *Conference on Language Modeling (COLM)*, 2024.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to reason with search for LLMs via reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jia Deng, Yimeng Chen, Xiaoqing Xiang, Ziyang Zeng, Shuo Tang, Wayne Xin Zhao, Feng Chang, Chuan Hao, Yuan Wei, Ran Tao, Bryan Dai, and Ji-Rong Wen. FORT-Searcher: Synthesizing shortcut-resistant search tasks for training deep search agents. *arXiv preprint arXiv:2606.12087*, 2026.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Unlocking long-horizon agentic search with large-scale end-to-end RL. In *International Conference on Learning Representations (ICLR)*, 2026.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

GLM-5 Team. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Klaudia Jaźwińska and Aisvarya Chandrasekar. AI search has a citation problem. Columbia Journalism Review, March 2025. URL https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, 2017. doi: 10.18653/v1/D17-1215.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Ö. Arık, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling (COLM)*, 2025.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations (ICLR)*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR)*, 2020.

Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, Qinyuan Cheng, Bo Wang, Yining Zheng, Yuxin Wang, Zhangyue Yin, and Xipeng Qiu. R3-RAG: Learning step-by-step reasoning and retrieval for LLMs via reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10491–10507, 2025. doi: 10.18653/v1/2025.findings-emnlp.554.

Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A Whac-A-Mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20071–20082, 2023. doi: 10.1109/CVPR52729.2023.01922.

Grace Liu, Yuxiao Qu, Jeff Schneider, Aarti Singh, and Aviral Kumar. CaRT: Teaching LLM agents to know when they know enough. *arXiv preprint arXiv:2510.08517*, 2025.

Nelson Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, 2023. doi: 10.18653/v1/2023.findings-emnlp.467.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations (ICLR)*, 2021.

Weimeng Luo. When should multi-round RAG stop? Structured stopping judgments and retrieval reduction in Search-R1. *arXiv preprint arXiv:2608.13237*, 2026.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019. doi: 10.18653/v1/P19-1334.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4249–4257, 2019. doi: 10.18653/v1/P19-1416.

Debayan Mukhopadhyay, Utshab Kumar Ghosh, and Shubham Chatterjee. Bridge evidence: Static retrieval utility does not predict causal utility in multi-step agentic search. *arXiv preprint arXiv:2607.15253*, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Thomas Schmied, Jörg Bornschein, Jordi Grau-Moya, Markus Wulfmeier, and Razvan Pascanu. LLMs are greedy agents: Effects of RL fine-tuning on decision-making abilities. In *International Conference on Learning Representations (ICLR)*, 2026.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Is multihop QA in DiRe condition? Measuring and reducing disconnected reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8846–8863, 2020. doi: 10.18653/v1/2020.emnlp-main.712.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

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

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. DeepTRACE: Auditing deep research AI systems for tracking reliability across citations and evidence. In *International Conference on Learning Representations (ICLR)*, 2026.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*, 2024.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. Process vs. outcome reward: Which is better for agentic RAG reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Wenlin Zhang, Kuicai Dong, Junyi Li, Yingyi Zhang, Xiaopeng Li, Pengyue Jia, Yi Wen, Derong Xu, Maolin Wang, Yichao Wang, Yong Liu, and Xiangyu Zhao. To search or not to search: Aligning the decision boundary of deep search agents via causal intervention. In *Proceedings of the ACM Web Conference 2026 (WWW)*, pp. 2049–2059. ACM, 2026a. doi: 10.1145/3774904.3792235.

Yulin Zhang, Yukun Huang, Sanxing Chen, Tianyi Lin, Ziang Yang, Xunjian Yin, and Bhuwan Dhingra. Lazy grounding: Attacking search agents with factual evidence. *arXiv preprint arXiv:2608.30303*, 2026b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

A THREE-SWITCH PAGES: CONSTRUCTION, CHECKS AND JUDGES

Questions and features. The parent questions are two-hop compositional questions from the 2WikiMultiHopQA development set with exactly two evidence triples, (X, r_1, K) and (K, r_2, W), where r_1 is father, mother, spouse or director and r_2 one of ten attributes. We excluded questions used in training or in the held-out split and questions that share X, K or W with our other materials, sorted the rest by identifier, shuffled them with a fixed seed and processed the first 100: 97 passed the mechanical checks, 66 passed two LLM reviews, and pre-screening and an audit removed 10 more, leaving 56 parent questions in 54 clusters. Of these, 35 had been used to evaluate the second set of training runs before this study. The question's entity X is replaced by invented names, as is a third entity used in some pages. The names were written by Claude Opus 5.5 without web access, under an instruction to avoid real names; 3 of the 112 names of question entities and 2 of the 56 names of third entities coincide with names of unrelated real entities, and 71 of the 112 aliases described below coincide with real names. A web search for further collisions was not completed, and the names of the held-out split were not checked against real names. The linked person K is real, W is K's real attribute and the original question's answer, and the background passages are real Wikipedia text, with the passage about K rewritten into a template sentence that keeps its fact. Because X is invented, a question has a correct answer only on pages that state the fact (448 of the 2,684 development pages). Each passage is shown with a question about each of two invented entities. In all main analyses the name slot holds an exact name, of the question's entity or of the other invented entity. Two further ways of referring to the subject are used only in Appendix B: an alias, kept only for questions where two reviewers judged the alias unambiguous, and a placeholder that names no entity ("a certain film"). When the candidate slot is filled, it holds W. The supporting fact is a sentence that links the subject to the answer through the asked relation. Relation-swapped pairs replace the relation word so that the sentence no longer answers the question, in one-hop and two-hop versions.

Checks. Every page is checked automatically for length and form, after the question-level screening above. Twenty pairs whose lengths differ by more than the tolerance are excluded from any estimate that uses both of their pages; none of them enters the main mention and candidate effects. Ten parent questions that may carry real-world cues form a preregistered sensitivity set: one whose question gives the real birth and death years of K's son, six whose invented film titles carry the real release year of a film K directed, and three whose invented titles are in the language of K's country or echo K's real work. Without them, the mention raises the fine-tuned agent's stopping rate by 49 [37, 62] percentage points and the candidate raises that of Qwen before fine-tuning by 33 [21, 46]; no main effect changes sign. The larger open models of Section 4.6 give an indirect check of the relation swaps, the edits that depend most on the reviewers: they stop much more often when the page states the asked relation than on the same page with the relation swapped (by 47–94 and 68–90 percentage points on one-hop and two-hop pairs), so the swap changes what a capable reader can conclude from the page.

Outcome labels. The next action is parsed from the model output. An LLM judge, blind to the condition, labels each final message as an answer, a decline or neither, and a second judge relabels all final messages. Whether an answer asserts the candidate is judged by the first judge, because models often answer in full sentences. Whether a model stops does not depend on the judges. The judges are Claude Opus 5.5 (first) and Claude Sonnet 5 (second), used through an agent interface with a fixed prompt template; this interface does not expose decoding parameters. Over all final messages they agree on 98% of labels (Cohen's $\kappa = 0.90$ [0.85, 0.93]), and they never disagree on whether a message is an answer or a decline; about half of the remaining disagreements (106 of 209) concern whether an answer addresses the question.

Splits. Questions were divided into a development split and a held-out split before the analyses reported in this paper, and the held-out files were sealed with recorded hashes. Before they were unsealed, two automated text searches for code identifiers also read these files, without displaying any of their content; no model had been run on them, and the analysis plan and code were already frozen. We record this as a deviation from the plan. The held-out split (83 parent questions in 78 clusters, 1,328 pages) was then unsealed at the authors' instruction and evaluated once, with the frozen plan and code.

B FULL RESULTS ON THREE-SWITCH PAGES

Log-odds of stopping. At the decision point we record the difference between the log-probabilities of the first token of a final answer and of a tool call; positive values favor stopping. Effects on this scale are averaged over the presence of the other feature, on pages without the fact.

Mention share and scale-and-shift fit. The mention share is the mention effect divided by the sum of the mention and candidate effects on the log-odds scale. A change in the overall willingness to stop that shifts or rescales all log-odds cannot change this share. We also regress the fine-tuned model's cell means on those before fine-tuning, $m_{\mathrm{ft}}(c) \approx a\, m_{\mathrm{pre}}(c) + b$ over page types c, and report the residual of the mention contrast (5.1 logits). This configuration was trained once, so this residual has no spread across seeds (an earlier run of the same recipe with an older script is in Table 2); fitted the same way, the three runs of the second set without counter-examples, which use slightly different data and are not seeds of the main agent, give 4.68, 4.21 and 3.70 logits.

Matching overall stopping rates. To ask whether a difference in an effect between two models or prompts reflects only a difference in how readily they stop, we shift the stop threshold of one until its overall stopping rate equals the other's, and recompute the effect.

Held-out confirmation. The held-out split has 83 parent questions in 78 clusters and 1,328 pages, disjoint from the development split. The plan fixed two primary comparisons on stopping rates: for the fine-tuned agent the mention effect exceeds the candidate effect (0.33 [0.24, 0.43]; development 0.42), and before fine-tuning the candidate effect exceeds the mention effect (0.23 [0.15, 0.31]; development 0.24). Both hold. Among the secondary comparisons, the mention share is higher for the fine-tuned agent by 0.74 [0.66, 0.83]; after matching stopping rates, the mention effect is lower with all counter-examples than without them by 0.12 [0.07, 0.18], above zero for all three seed pairs; and the fine-tuned agent answers on 97.9% [96.0, 99.4] of pages that state the fact. The fine-tuned agent's candidate effect is larger on the held-out pages than on the development pages (12.5 against 5.3 points); this comparison was not preregistered.

Placeholders and aliases. When shown only a page that refers to the subject by a placeholder ("a certain film"), all four models we asked, with and without fine-tuning, answered that they could not tell which entity the page is about in 98–100% of cases, and the two LLM judges of Appendix A, labeling blind, could not tell for any of the 112 distinct page texts. We therefore report comparisons that involve placeholders only descriptively. With counter-examples, the effect of an exact name relative to a placeholder fell, and 59% of the fall remained after matching stopping rates. The fine-tuned agent stops on 55% of pages with the exact name, 19% with an alias and 8% with a placeholder or another entity; before fine-tuning the rates are 86%, 87%, 78% and 81%. Counter-examples lower only the first two (by 0.22 and 0.08). Because many aliases coincide with names of real entities (Appendix A), we do not interpret the alias rates further.

Instructions. The three instructions, appended verbatim after a blank line at the end of the system prompt, were: (i) "Only give a Final Answer if the retrieved text explicitly states the fact needed to answer the question. If the text is about the entity the question asks about but does not state that fact, use the search tool again." (ii) "Before giving a Final Answer, check that the retrieved passages contain the specific fact the question asks for. A passage that only mentions or describes the entity in the question is not enough; in that case, search again." (iii) "Only give a Final Answer when the retrieved text contains enough information to answer the question. Otherwise, use the search tool again." Instruction (i) was run on Qwen before fine-tuning, the main agent and the six runs of the second set with no or all counter-examples; (iii) on Qwen before fine-tuning, the main agent and the three runs without counter-examples; (ii) on the main agent and the three runs without counter-examples. All 17 runs completed on 1,120 pages each. The rule for reading them was fixed before the runs: an instruction changes what a model stops on only if its mention effect falls after matching overall stopping rates, with an interval that excludes zero (for the three seeds, pooled, and with at least two of three seeds agreeing in sign). By this rule no instruction changed what any fine-tuned run stops on; none of the 17 selective changes has an interval entirely below zero. The only selective changes we found were before fine-tuning, in the two-hop effect of the fact, under instructions (i) ($+0.165$ [0.07, 0.26]) and (iii) ($+0.191$ [0.08, 0.31]). On the development pages, the

matched effect of instruction (i) on the mention cue of the runs without counter-examples was 0.23 times the effect of the counter-examples; the instruction's effect has an interval that includes zero, and the ratio was not fixed in advance and has no interval. Results are in Table 3.

Table 3: All 17 instruction runs on the 1,120 development pages. Overall stopping rate without and with the instruction; change in the mention effect on stopping (percentage points), raw and after matching the overall stopping rate (selective; 95% cluster-bootstrap interval).

Instruction	Model	Stop, without	Stop, with	Raw	Selective
(i) names the situation	Before fine-tuning	0.88	0.42	−9	−3 [−9, 4]
	Main agent	0.43	0.40	−14	−4 [−9, 0]
	No counter-ex., s42	0.48	0.49	−2	−4 [−10, 2]
	No counter-ex., s43	0.46	0.45	−4	−1 [−7, 5]
	No counter-ex., s44	0.41	0.43	−1	−3 [−7, 1]
	All counter-ex., s42	0.40	0.40	−1	0 [−4, 3]
	All counter-ex., s43	0.40	0.38	−4	−4 [−10, 1]
	All counter-ex., s44	0.36	0.36	0	0 [−4, 5]
(ii) reworded	Main agent	0.43	0.35	−19	−2 [−6, 3]
	No counter-ex., s42	0.48	0.43	−9	−2 [−6, 3]
	No counter-ex., s43	0.46	0.41	−10	−1 [−5, 2]
	No counter-ex., s44	0.41	0.40	−2	2 [−2, 6]
(iii) generic	Before fine-tuning	0.88	0.49	−2	4 [−1, 9]
	Main agent	0.43	0.42	−4	0 [−5, 5]
	No counter-ex., s42	0.48	0.50	1	−4 [−9, 3]
	No counter-ex., s43	0.46	0.47	4	2 [−3, 7]
	No counter-ex., s44	0.41	0.44	7	4 [0, 7]

Asking versus acting. The plain-format question was "Based only on the retrieved text above, does it state the specific fact needed to answer the question? Answer Yes or No.", and we compare the log-probabilities of "Yes" and "No" at the first answer token. The plan made the question in the agent's own input format primary and prescribed the plain prompt if that format gave a valid answer on fewer than 95% of pages, as it did for the fine-tuned agent (78%); the verdict therefore uses the plain prompt. The plan fixed three conditions before the run: the fine-tuned agent's accuracy and its balanced accuracy in judging whether the fact is stated must each be at least 0.8 (as point estimates), and the lower end of the interval for the difference between the mention's effect on stopping and on saying yes must be above zero. All three hold (accuracy 0.900 [0.872, 0.926], balanced accuracy 0.825 [0.781, 0.868], difference 0.444 [0.338, 0.556]). On the 224 pages, the agent says the fact is missing on 109 of 112 pages without the candidate and 105 of 112 with it. A check suggested after seeing the data, and not in the plan, indicates that the agent's judgment follows a fact-like candidate more than its action does. Results are in Table 4.

Table 4: Asking versus acting on the development pages. Plain and balanced accuracy of the yes-or-no judgment in the plain prompt, over all pages; mention effects (percentage points) on stopping and on saying "yes" on pages without the fact, averaged over the candidate; and the share of valid yes-or-no answers when the question was asked in the agent's own input format.

Model	Plain acc.	Balanced acc.	Stop	"Yes"	Difference	Valid, agent format
Before fine-tuning	0.87	0.72	6	−2	8 [3, 14]	0.57
Main agent	0.90	0.82	47	3	44 [34, 56]	0.78
No counter-ex., s42	0.87	0.83	45	4	41 [31, 51]	0.69
All counter-ex., s42	0.87	0.79	20	1	19 [13, 27]	0.86

Forced choice in the agent's input format. This exploratory check appends the same question to the agent's own input and compares the summed probabilities of 6 single-token forms of "Yes" and 7 of "No" at the first answer token. Other forms, with leading punctuation or symbols, were not counted. On the 224 pages that mention the subject without the fact, "No" wins on 195 (87%) for the

fine-tuned agent and on 211 (94%) before fine-tuning. The fine-tuned agent stops on 121 of these pages (120 answers and one decline), 92 of them pages it judged to lack the fact; before fine-tuning the model stops on 192, and on pages without the candidate 65 of its 81 stops are declines. Balanced accuracy in judging whether the fact is stated is 0.885 [0.848, 0.920] after and 0.747 [0.715, 0.779] before fine-tuning; the lower value before fine-tuning comes mainly from two-hop pages with the fact, on which the model puts little probability on either answer (median about 0.12) and prefers to act. The mention raises the share of "Yes" judgments of the fine-tuned agent by 0.10 [0.05, 0.16], and its effect on stopping exceeds its effect on the judgment by 0.37 [0.26, 0.48], against 0.08 [0.03, 0.13] before fine-tuning. Before fine-tuning, the candidate moves stopping more than the judgment (by 0.155 [0.028, 0.280]), whereas for the fine-tuned agent it moves both about equally (-0.005 [-0.056, 0.042]). Because the model before fine-tuning puts little probability on either answer in this format (median 0.27 on the pages where it answered with the candidate), we do not read its judgments page by page.

Relation-swapped pairs. Stating the asked relation raises the fine-tuned agent's stopping rate by 0.52 [0.39, 0.64] on one-hop and 0.52 [0.40, 0.65] on two-hop pairs, and by 0.42 to 0.58 for the six runs of the second set with no or all counter-examples. Before fine-tuning the stopping rate on these pages is at ceiling (0.99 and 0.98), and the answer rate rises by 0.31 [0.20, 0.44] and 0.35 [0.23, 0.46]. After matching stopping rates, counter-examples change the effect of support by less than 0.10, although the one-hop estimate is negative for all three seeds.

Table 5: Outcomes of Qwen2.5-7B-Instruct before and after fine-tuning on three-switch pages with the exact name (development split; counts out of 112 pages per row). *Other* counts final messages the judge labels neither an answer nor a decline; *Candidate* counts final messages that the first judge labels as asserting the candidate; before fine-tuning these include some messages counted under *Other* (on pages without the candidate, the string comes from the model's own knowledge).

Model	Page type	Answer	Decline	Other	Search	Candidate
Before	neither	12	56	1	43	3
	candidate	66	35	8	3	71
	mention	13	65	3	31	7
	mention + candidate	84	23	4	1	85
	fact, one hop	112	0	0	0	112
	fact, two hops	106	4	2	0	108
	relation swapped, one hop	77	31	1	3	74
	relation swapped, two hops	66	37	5	4	65
After	neither	5	0	0	107	0
	candidate	12	0	0	100	12
	mention	57	1	0	54	3
	mention + candidate	63	0	0	49	59
	fact, one hop	112	0	0	0	110
	fact, two hops	110	0	0	2	107
	relation swapped, one hop	55	0	0	57	55
	relation swapped, two hops	50	0	0	62	48

Log-odds effects, including the comparisons between ways of referring to the subject, are in Table 6.

Reasoning before acting. This exploratory run starts the model's reply with "Thought:", so that it writes its reasoning before an action, on the same 1,120 development pages as the instruction runs, with greedy decoding and up to 512 new tokens; the reply is read up to the first "Observation". The reading rule was fixed before the run: a cue counts as reduced only if the interval of its change excludes zero. Both models gave a valid action on every page, although the fine-tuned agent was never trained in this format. With the thought, the main agent's mention effect on stopping fell from 0.47 to 0.11 (change -0.36 [-0.48, -0.25]), the effect of the fact on relation-swapped pairs, averaged over one and two hops, rose from 0.52 to 0.66 ($+0.14$ [0.03, 0.25]), and its overall stopping rate fell from 0.43 to 0.28. Before fine-tuning, the candidate effect did not fall (0.29 to 0.43; change $+0.14$ [-0.02, 0.30]); the effect of the fact rose from 0.03, with stopping at ceiling, to 0.42, and the overall stopping rate fell from 0.88 to 0.42. The other two effects changed by less than 0.06, with intervals that include zero. The thought is prefilled rather than chosen by the model, each page was

run once, and pages were decoded in batches rather than one at a time, which can change greedy outputs slightly.

Table 6: Effects on the log-odds of stopping (logits) on development pages, with 95% cluster-bootstrap intervals. Mention and candidate effects are on pages without the fact, averaged over the other feature; the lower rows compare the exact name with an alias (the 52 questions with an unambiguous alias) and with a placeholder that does not name the subject, and compare that placeholder with a page that names the other entity.

Effect	Before fine-tuning	After fine-tuning
Mention (exact name)	0.21 [−0.08, 0.47]	5.12 [4.48, 5.77]
Candidate	4.30 [3.28, 5.38]	0.60 [0.28, 0.94]
Mention share	0.05	0.89
Fact, one hop	4.57 [3.81, 5.36]	5.42 [4.58, 6.26]
Fact, two hops	2.12 [1.61, 2.63]	5.48 [4.59, 6.38]
Mention by alias	0.20 [−0.02, 0.42]	2.46 [1.99, 2.97]
Exact name minus placeholder	0.47 [0.14, 0.79]	5.03 [4.38, 5.68]
Placeholder minus other entity	−0.27 [−0.46, −0.09]	0.09 [0.02, 0.16]
Exact name minus alias	0.09 [−0.11, 0.28]	2.76 [2.24, 3.28]

C MODELS WITHOUT OUR FINE-TUNING

Protocol. All models use the same ReAct-style prompt with a single search tool (Yao et al., 2023), except the larger Qwen models below, which were run under an earlier version of it, and the models of Section 4.6, which received each page through their providers' APIs (below). The observation returned by the first search is replaced by the edited page, the next action is decoded greedily, and a fixed parser reads it. An answer uses the candidate if the final answer contains the candidate string after normalization. None of the 600 records with and without a candidate in Table 1 failed to parse. For MuSiQue, the 50 triples were selected from 80 that passed construction, built from about 1,252 two-hop questions. The system prompt names one tool, `search(query)`, and asks for either "Action: <tool_name>" with "Action Input: <input>" or "Final Answer: <answer>"; the user turn gives the question, states that a search has already been run, shows the tool result and repeats the two response shapes (full text in Appendix G; SHA-256 hashes `26daada9...` and `830b2414...`). On MuSiQue, the page without a candidate holds four distractor passages; the page with an unsupported candidate holds the second-hop supporting passage, which contains the answer, and three distractors, but not the first-hop passage that links it to the question; the supported page holds both supporting passages and two distractors. The template items are generated from 8 bridge templates with invented entities, in the style of HotpotQA bridge questions (6 or 7 triples per template); there the candidate appears in a sentence that does not connect it to the question's entity, and the supported page adds the linking facts and an answer sentence. Pages of a triple were kept only if their token lengths differ by at most a factor of 1.35. In both sets the candidate is the answer of the question (50 of 50), so the unsupported page contains the correct answer without the facts that support it. A supporting sentence raises answering with the candidate further: for Qwen by 68 percentage points on the template items but only 12 on MuSiQue, where an unsupported candidate is taken almost as often as a supported one; for Mistral by 42 and 14 ($p = 0.09$ on MuSiQue) and for Gemma by 90 and 8 ($p = 0.34$).

2WikiMultiHopQA. On a filtered subset of 2WikiMultiHopQA (Ho et al., 2020), adding an unsupported candidate raises answering with it by 0.36 for Qwen ($p = 7.6\times10^{-6}$), 0.28 for Mistral ($p = 1.3\times10^{-3}$) and 0.10 for Gemma ($p = 0.063$). The subset keeps only questions whose pages pass the construction checks, about 1.2% of the questions we tried, so it cannot estimate how common the situation is.

Sampled decoding. On 50 candidate pages under a less restrictive version of the prompt, sampling five continuations per page (temperature 0.7, top-p 0.95) gives a mean increase in the probability of answering with the candidate of 0.76 [0.68, 0.84] for Qwen, against 0.84 [0.74, 0.94] with greedy decoding under the same prompt.

Larger Qwen models. On the MuSiQue pages of Table 1, adding the unsupported candidate raises answering with it by 0.30 for Qwen2.5-14B-Instruct, 0.22 for Qwen2.5-32B-Instruct and 0.40 for Qwen3-32B (Qwen Team, 2025); all three differences are significant (exact McNemar $p = 6.1\times10^{-5}$, 9.8×10^{-4} and 1.1×10^{-5}). Without the candidate, these models give it as their answer in at most 10% of cases. On the 50 template items of Table 1, Qwen3-32B answers with the candidate on every page that contains it and on none without it ($p = 1.8\times10^{-15}$). Qwen3-32B is about the size of Qwen2.5-32B-Instruct but comes from a different training recipe. These runs used an earlier version of the system prompt, and the size of the effect depends strongly on the prompt (for Qwen2.5-7B-Instruct on the template items, 0.32 under the prompt of Table 1 and 0.84 under the less restrictive prompt of the sampled-decoding check above), so we compare only its direction.

Code search and SQL. In code search, the tool returns a legacy file whose name matches the query instead of the file on the production path; Qwen answers with it in 60% of cases, against 0% with no such file ($p = 2.9\times10^{-11}$), and answers correctly in 93% when the canonical file is shown. In SQL, a prefilled query computes the wrong metric and its top row looks like an answer; Qwen commits to that row in 78% of cases, against 0% without it, and answers correctly in 100% when the right query is shown. In a single-prompt version of both tasks the rates drop (60% to 5% and 78% to 2%); because that version also changes the instructions, the available actions and the parser, we cannot attribute the drop to the tool observation alone. Tools, parsers and baselines differ from question answering, so we compare directions, not magnitudes.

Unedited states before fine-tuning. We ran Qwen2.5-7B-Instruct on 500 HotpotQA bridge questions and took the state right after its first search. It answered in 470 states. By exact match, 373 answers differ from the gold answer; in 80 of them one string contains the other. Of the other 293, 219 (75%) are strings from the retrieved text, 121 while the gold answer is not in the text and 98 while it is also there; 51 appear nowhere in the text while the gold answer does not either, and 23 appear nowhere although the gold answer does. This breakdown is descriptive; it does not show why the model chose these answers.

Llama-3.1-8B-Instruct. Under our prompt Llama searches after every template page and after 98% of supported MuSiQue pages, so this test is uninformative for Llama. The edits do move its log-odds of stopping: on the template items they rise from -21.2 with no candidate to -19.9 with an unsupported candidate and -14.9 with a supported one (Wilcoxon $p \approx 4\times10^{-10}$ and 2×10^{-15}). When a supported page also states the answer explicitly, Llama often copies the candidate into its next search query.

Larger open models. The four models (API names deepseek-v4-pro, deepseek-flash, glm-5.3 and glm-5.3-flash) saw 1,120 of the 2,684 development pages: the 448 pages with the exact name that vary the mention and the candidate, 224 placeholder pages, and the 448 pages that state the fact or swap the relation. Each page came as the result of a search tool call, through each provider's official interface and native tool format, on 25 September 2026, at temperature 0 with up to 16,000 output tokens, one sample per page and the default thinking mode on. The reasoning text was archived but not used for labeling or shown to the judges. The interfaces report no snapshot dates; DeepSeek's system fingerprint stayed the same throughout. A tool call counts as a search and a text reply as a stop; replies were labeled by the two judges of Appendix A, blind to the condition, who agreed on 777 of 777 labels, and the mechanical check of whether an answer contains the candidate agreed with the judges. Under a reworded neutral system prompt, actions agreed on 91–97% of pages. One provider's content filter blocked the outputs for one parent question (17 pages in the main run and 18 under the reworded prompt), which we exclude under an amendment written before the analysis; the conclusions do not change without it. The classification rule, fixed before the run, calls a model like the fine-tuned agent if its mention effect on stopping is positive, with an interval that excludes zero, and larger than its candidate effect; like Qwen before fine-tuning if its candidate effect is positive and larger than its mention effect; and like neither otherwise. By this rule no model resembles the fine-tuned agent, also after excluding the parent questions with possible real-world cues (Appendix A). GLM-5.3-Flash meets the rule's criterion for resembling the model before fine-tuning on the full set (candidate effect 2.5 against mention effect 1.6 percentage points) but not after that exclusion, and all its mention and candidate effects are below 3 points. Per-model stopping rates are in Table 7.

Table 7: Stopping rates (%) of the four open models on the development pages they saw, by page type: without the fact, pages with the mention only (M), the mention and a candidate (M+C) or a candidate only (C); pages that state the fact; and relation-swapped pages, on one- and two-hop pairs. Rates are averaged within parent questions, then over clusters.

	No fact			Fact		Relation swapped	
Model	M	M+C	C	1 hop	2 hops	1 hop	2 hops
DeepSeek-V4-Pro	0	3.2	0	96.8	92.6	2.8	2.3
deepseek-flash	0	0.9	0	85.2	88.9	4.6	1.9
GLM-5.3	0	6.9	0	47.6	91.5	0.9	7.1
GLM-5.3-Flash	0	4.2	0.9	94.3	98.1	25.0	30.2

Thinking turned off. Of the four models, only the two DeepSeek models accept a request to turn thinking off; GLM-5.3 rejects it, and GLM-5.3-Flash ignores it and still reasons. This arm followed an amendment written before it was run, and all other parameters were those of the first run. Table 8 gives the five preregistered quantities with and without thinking. Counting only answers that contain the candidate string, the candidate effect without thinking is 0.19 and 0.51 and the mention effect 0.12 and 0.10, against at most 0.02 with thinking. On one-hop relation-swapped pages the two models answered with the candidate on 84 and 108 of 112 pages (3 and 5 with thinking), and on two-hop pages on 48 and 69. We split stops into answers and declines by their wording with a fixed regular expression, without the judges, so this split is descriptive: without thinking, 22 of 22 and 49 of 50 stops on pages with only a mention were declines, against 28 and 73 declines on pages with neither the mention nor the candidate. With thinking on, DeepSeek ignores the temperature parameter, so the requested temperature of 0 took effect only in this arm. Under a further amendment, both models were rerun without thinking at temperature 1.0, the interface's default, on the one-hop relation-swapped pages, the pages with only a mention and the pages with only the candidate (112 each). The three preregistered rates changed by at most 0.03, with intervals that include zero; for example, the two models answered with the candidate on 86 and 109 of the 112 one-hop relation-swapped pages, against 84 and 108 at temperature 0.

Table 8: The two DeepSeek models with thinking turned off and on: stopping rate on pages with only a mention, and effects on stopping (percentage points) of the mention, the candidate and the fact on one- and two-hop relation-swapped pairs. Differences are paired over question clusters, with 95% cluster-bootstrap intervals.

	DeepSeek-V4-Pro			deepseek-flash		
	Off	On	Off − on	Off	On	Off − on
Stopping, mention only	20.4	0.0	20.4 [11.1, 30.6]	46.3	0.0	46.3 [33.3, 58.3]
Mention effect	6.5	1.6	4.9 [−2.3, 12.0]	−9.3	0.5	−9.7 [−18.5, −1.2]
Candidate effect	10.2	1.6	8.6 [0.5, 16.7]	21.3	0.5	20.8 [12.3, 29.9]
Fact, one hop	25.9	94.0	−68.1 [−77.3, −58.3]	3.7	80.6	−76.9 [−85.2, −67.6]
Fact, two hops	34.7	90.3	−55.6 [−70.8, −38.9]	6.5	87.0	−80.6 [−89.8, −70.4]

Other families with reasoning turned off. Under an amendment written before these runs, four models from other families ran on the same 1,120 pages with reasoning turned off: Kimi K3, MiniMax M3, Mistral Medium 3.5 and Qwen3.8 Flash (the flagship Qwen model does not allow reasoning to be turned off). They were accessed through one routing interface with the first-party provider fixed and no fallback; no reply contained reasoning text or reasoning tokens. Other settings match the first run, except that Kimi K3's interface rejects a temperature, so its default was used. All quantities are mechanical, without the judges: a text reply is a stop, and an answer uses the candidate if it contains the candidate string. By the rule fixed before the first run (Table 9), Kimi K3, MiniMax M3 and Mistral Medium 3.5 resemble Qwen before fine-tuning, each with an interval for the difference between the candidate and mention effects that excludes zero, and Qwen3.8 Flash, which almost never stops without the fact, resembles neither. Kimi K3 and MiniMax M3 barely distinguished pages that state the fact from near misses, and they stopped on 36 and 42 of the 112 pages with only a mention; by their wording, many of these stops answer the second hop from the model's own

knowledge of the linked person instead of declining. MiniMax M3 also ran with its default reasoning (55 reasoning tokens per reply on average): its answers with the candidate on one-hop near misses fell from 108 to 47 of 112 (-0.56 [-0.66, -0.46]), and its stops on pages with only a mention from 42 to 4. A matching run of Qwen3.8 Flash with reasoning did not finish because of rate limits and is not analyzed. Models from OpenAI, Anthropic and Google were not available to us through this interface.

Table 9: Models with reasoning turned off (and MiniMax M3 with its default reasoning) on the development pages: answers with the candidate on relation-swapped pages (near misses) and stops on pages with only a mention, out of 112 pages each; effects on stopping (percentage points) of the mention, the candidate and the fact (one-hop relation-swapped pairs); and the model before or after fine-tuning that each resembles under the rule fixed in advance. The effects of the two DeepSeek rows repeat Table 8.

	Near misses		Mention	Effect on stopping			
Model	1 hop	2 hops	only	Mention	Cand.	Fact	Resembles
Kimi K3	108	97	36	12.7	44.2	2.8	before
MiniMax M3	108	91	42	6.0	42.6	3.2	before
Mistral Medium 3.5	40	27	0	2.3	15.3	63.9	before
Qwen3.8 Flash	13	4	0	3.2	3.2	62.0	neither
MiniMax M3, reasoning on	47	54	4	11.3	17.8	54.6	before
DeepSeek-V4-Pro	84	48	22	6.5	10.2	25.9	before
deepseek-flash	108	69	50	-9.3	21.3	3.7	before

D REAL SEARCH EPISODES

First decisions. The episodes come from the first step of an earlier evaluation loop on 600 questions: 300 training questions of MuSiQue-Ans (228 two-hop, the rest three- or four-hop; none used to fine-tune our agents) and 300 development questions of 2WikiMultiHopQA (157 comparison, 99 bridge-comparison and 44 inference questions). Comparison questions name two entities and have two possible answers, so some correct answers in states with incomplete support can be guesses. From the second step on, that loop presented results in a format the agents never saw in training, so we use only the first step, whose inputs are identical to those of the corrected loop (checked byte for byte on 800 inputs). In the corrected loop, many episodes repeated the same query until the search budget ran out (in 32–44 of 200 development episodes per model the same query was sent four times), so we did not measure end-to-end accuracy. A page counts as being about the named entity if one retrieved passage is titled with the entity and mentions it in its body; the entity is the title of the first hop's gold supporting page. Support is complete if all gold supporting passages, counted up to near-duplicates, are in the first results. We analyze the 305 questions whose wording names this entity and whose entity is not itself the bridge or the answer string: 35 have the page about the named entity and complete support, 199 that page but incomplete support, 3 complete support without that page, and 68 neither. The regression includes both predictors and their interaction. With only 3 questions in the third group, the interaction and the coefficient for complete support have wide intervals (in a model without the interaction, complete support raises the log-odds by 1.85 to 2.51 across runs, with intervals that exclude zero), while the coefficient for the named page compares the 199 questions with the 68. Standard errors are clustered by question cluster.

An early stop in a real episode

Question (2WikiMultiHopQA): Do both The Badlanders and The Man From Monterey films have the directors from the same country?

First results: the pages on *The Man from Monterey*, *Shake! Otis at Monterey*, *The Badlanders* and *Monterey Media*; not the pages on the two directors. The page about the named entity begins: "The Badlanders (1958) is a western caper film directed by Delmer Daves and starring Alan Ladd and Ernest Borgnine. Based on the 1949 novel 'The Asphalt Jungle' by W. R. Burnett, the story was given an 1898 setting by screenwriter Richard Collins."

Agent, first decision: Final Answer: No **Correct answer:** yes (both directors are American)

Effect of the page about the named entity. Logistic coefficients: main agent 1.03 $[0.44, 1.61]$; second set without counter-examples, seed 42 0.40 $[-0.15, 0.95]$, seed 44 0.68 $[0.07, 1.29]$, seed 43 0.17 $[0.04, 0.31]$ as a change in probability from a linear model because the logistic fit separates; second set with all counter-examples, seed 42 0.16 $[-0.42, 0.73]$. Per-group counts are in Table 10.

Table 10: First decisions on the 305 analyzed questions of the 600: stops and correct answers per group, by whether the first results include the page about the named entity (named; other: not) and all supporting passages (full; partial: not all). Group sizes: 35, 199, 3 and 68. Runs: the main agent and the second set without counter-examples (seeds 42–44) and with all of them (seed 42).

	Named, full		Named, partial		Other, full		Other, partial	
Run	Stops	Correct	Stops	Correct	Stops	Correct	Stops	Correct
Main agent	31	21	107	35	2	2	20	9
No counter-ex., s42	32	21	83	30	2	1	22	9
No counter-ex., s43	35	24	119	43	3	1	29	9
No counter-ex., s44	31	20	79	28	2	2	17	8
All counter-ex., s42	26	15	59	21	2	1	18	8

Questions the agents cannot know. A separate set of 200 questions, none used to fine-tune our agents (100 two-hop MuSiQue-Ans training questions and 100 compositional 2WikiMultiHopQA development questions, disjoint from the 600) was drawn from those that the run of the second set without counter-examples (seed 42) answers wrongly when shown only the gold page for the first hop and forced to answer. The same 118 of them reach a state with the page about the named entity but incomplete support for every run, since the first observation does not depend on the model. In these states the main agent and the two runs of the second set with seed 42, without and with counter-examples, stopped in 52%, 42% and 35% of cases, against 54%, 42% and 30% in the same states on the 600 questions, and answered correctly in none of their 61, 50 and 41 answers (Wilson upper bounds 5.9%, 7.1% and 8.6%). Only the run without counter-examples was used to select these questions. The main agent's own probe, run the same way, marks 197 of the 200 as questions it answers wrongly from the first-hop page alone; on the 117 such questions among the 118, it stopped in 61 and answered none correctly. Outside these states, some first-step answers on the same questions were correct (16 of 29 for the main agent and 15 of 29 for the selecting run), so the zeros are not an artifact of scoring.

Later states. We also re-presented 412 later states from the trajectories of the second set's run without counter-examples (seed 42), in which the entity needed for the next hop had been retrieved, in the fine-tuned agents' input format. The main agent stopped in 30% of them, with 21% of its answers correct; the two runs of the second set stopped in 16% and 13%, with 11% and 21% correct. Qwen before fine-tuning, which was not trained on this format, stopped in 99% of states, declining in at least 51%. Split by whether the second-hop passage had been retrieved (320 and 92 states; only 2 states had complete support), the main agent stopped in 27% and 38% of states, with 11% and 46% of its answers correct. When the fine-tuned agents searched, their queries named the bridge entity in 36–37% of cases and added the relation in 1 of 995; in the training data every search target names the bridge entity, taken from the dataset's annotations (Appendix E), which the agent at test time does not have. We have no results for Mistral or Gemma on these states. Counts are in Table 11.

Table 11: Later states (412): stops and correct answers among stops, overall and split by whether the passage for the second hop had already been retrieved (320 states without it, 92 with it); –: not split for the model before fine-tuning.

	All (412)		2nd hop missing (320)		2nd hop present (92)	
Model	Stops	Correct	Stops	Correct	Stops	Correct
Before fine-tuning	408	16	–	–	–	–
Main agent	122	26	87	10	35	16
No counter-examples, seed 42	66	7	56	4	10	3
All counter-examples, seed 42	53	11	40	4	13	7

E FINE-TUNING DATA AND RUNS

Training data. The main agent's examples were built from 200 MuSiQue and 44 2WikiMultiHopQA questions. For each question we built three observations: one containing the passages that state the answer (target: the answer), one containing an answer-like string without support, and one without it (both with target: a search whose query is the bridge entity from the dataset's annotations). Questions were removed when the answer could not be extracted from the second-hop passage (MuSiQue, 422), when support could not be resolved (2WikiMultiHopQA, 4,646), when the length ratio of the observations was out of range (258 and 495), for other reasons (23 in total), or when they overlapped the held-out questions (38 and 50). This gave 732 examples, of which a frozen subset of 706, from 239 of the 244 questions, was used for training in every run except the earlier run described below; the other 26 exceed 1,024 tokens with their own target or with the permuted target of a control run. None is a decline, and in 4 observations meant to lack the answer, the answer string does appear. The data of the second set without counter-examples differ from the main agent's only in one distractor passage in 144 of the 706 training examples, with identical targets; its seed-42 run uses the same trainer, initialization and seed as the main agent. The counter-example for a question is its first-hop page with every sentence that names the next entity or the answer removed; it has to keep at least 40 characters and pass two LLM reviews checking that it no longer states the link. In each version of the second set, one distractor passage in each of the 144 eligible search examples is replaced: by the counter-example in the first k examples of a fixed random order ($k = 0$, 16, 53 and 144, nested across versions) and by an unused passage of the same question otherwise. All runs use low-rank adapters (rank 16, $\alpha = 32$, dropout 0.05), 3 epochs, learning rate $5{\times}10^{-5}$ with 10% warm-up, batch size 2 with 4 accumulation steps and gradient clipping at 0.3. The main agent and the second set are trained on the 706 examples (264 optimizer steps) with the loss on target tokens only; the maximum length is 1,024 tokens for the main agent and 1,186 for the second set. The other run of the main agent's recipe was trained earlier with an older training script on the same file of 732 examples, with a maximum length of 1,024 tokens; it took 237 optimizer steps, and the script did not record how many of the examples it kept. Retraining one run of the second set (0 of 144, seed 42) with its original configuration reproduced its adapter bit for bit, which shows that training is deterministic; it is not an independent replication. Table 12 lists every run.

Table 12: Every fine-tuning run evaluated on the development pages, except a control trained on permuted targets, which never stops: counter-examples among the 144 eligible search examples, seeds, and the mention and candidate effects on stopping (percentage points).

Run	Counter-examples	Seeds	Mention	Candidate
Main agent	main agent's data	42	47.0	5.3
Main agent's recipe, earlier run	main agent's data	42	40.5	14.1
Second set	0 of 144	42 / 43 / 44	44.9 / 51.6 / 34.0	11.6 / 12.7 / 17.4
Second set	16 of 144	42	39.4	15.3
Second set	53 of 144	42	30.3	18.8
Second set	144 of 144	42 / 43 / 44	20.4 / 25.9 / 15.5	15.3 / 19.0 / 19.2
Mistral-7B-Instruct-v0.3	main agent's data	42	35.0	−0.2
Mistral-7B-Instruct-v0.3	144 of 144	42	21.5	0.2

Mistral-7B-Instruct-v0.3. Under an amendment written before training, we fine-tuned Mistral-7B-Instruct-v0.3 with the main agent's recipe, once on the main agent's data and once on the second set's data with all counter-examples, both with seed 42, and evaluated both, and the model before fine-tuning, once on the 1,120 development pages of the instruction runs. The maximum length was raised to 1,152 and 1,280 tokens so that all 706 examples fit, both runs took 264 steps, and pages were decoded in batches. All three models gave a valid action on every page. After fine-tuning on the main agent's data, the mention raised the stopping rate by 0.350 [0.243, 0.463] and the candidate by -0.002 [-0.037, 0.032]; before fine-tuning, the candidate raised it by 0.403 [0.287, 0.519] and the mention by 0.037 [-0.005, 0.079]. The difference between the two effects is 0.352 [0.241, 0.472] after and 0.366 [0.241, 0.491] in the other direction before fine-tuning. With all counter-examples the mention effect was 0.215 [0.127, 0.315], lower by 0.122 [0.021, 0.224] after matching overall stopping rates. Unlike the fine-tuned Qwen agent, fine-tuned Mistral barely distinguished pages that state the fact from near misses (effect of the fact on stopping 0.056 [-0.056, 0.167] on one-hop and 0.120 on two-hop pairs; 0.259 and 0.343 with all counter-examples), although it stopped on 84% of the pages that state the fact. Each configuration was trained once, so the intervals reflect only the sampling of questions.

Public training data of search agents. We applied the coupling rule to three public training sets of multi-hop search agents whose decision states list the retrieved passages with their titles: the process-supervision preference pairs of RAG-ProGuide (Zhang et al., 2025), the decomposed-query trajectories of RQ-RAG (Chan et al., 2024) and the cold-start trajectories of R3-RAG (Li et al., 2025). The rules were fixed before any data were downloaded; the last two datasets were added later, each under an amendment written before its analysis, and two parsing errors found in our own checks were corrected, also under a written amendment, before the numbers reported here. We keep states that already have retrieved text and whose question matches a bridge question of HotpotQA, 2WikiMultiHopQA or MuSiQue with a known order of hops; 88%, 100% and 99.7% of states matched. In these data a page about the named entity barely predicts stopping, and about half of the examples whose target is a search come after the first-hop page has been retrieved but not the second (Table 13). The 90% cluster-bootstrap intervals fixed in the plan are within ± 0.03 for the AUROC and ± 0.06 for the shares. Five further datasets could not be scored, because their observations lack page titles or per-step passages or because fewer than half of their states matched. Three caveats apply: the observations are the retrieved passages as recorded in each dataset; RQ-RAG probably retrieves from a closed pool of passages per question; and the trajectories of RAG-ProGuide and R3-RAG were generated by LLMs and filtered by the correctness of the answer. Support is judged by page titles, which can count a distractor that shares a title with a supporting page; scored by titles as well, our 732 source examples give 10 of the 444 search examples that pass the same filters in this state.

Table 13: Coupling between a page about the named entity and stopping in training data, among states that already have retrieved text: the share of states with that page among states whose target is to stop and among states whose target is a search, the AUROC of that page as a predictor of stopping (one half plus half the difference of the two shares), and the examples whose target is a search although the first-hop page has been retrieved and the second has not. When the two shares are close, the page barely predicts stopping. The public sets are scored by page titles; the preference data of Zhang et al. (2026a), whose pages have no titles, by mentions; our data by passages, with the AUROC on the examples whose question names the entity and the last column over all search examples.

			Page present			Search after
Data	Kind	States (questions)	stop	search	AUROC	first hop only
Main agent's data	single decisions	461	1.00	0.04	0.98	0 of 473
RAG-ProGuide	preference pairs	1,373 (751)	0.59	0.55	0.52	153 of 315
RQ-RAG	trajectories	28,536 (11,271)	0.92	0.83	0.54	8,261 of 17,265
R3-RAG	trajectories	73,820 (29,742)	0.68	0.55	0.56	21,863 of 44,078
Zhang et al. (2026a)	preference pairs	3,181	0.75	0.64	0.56	76 of 444

Coupling statistics. For each training example we record whether the observation includes a page about the question's entity, whether it states the fact, and whether the target is an answer. Coupling is

the AUROC of a feature as a predictor of the target. The primary rule, fixed in the preregistration of the second set of runs, counts a passage titled with the entity; under it the coupling is 0.98, 0.98 and 0.76 for the main agent's data and the second set without and with all counter-examples (computed on the 461 examples whose question names the entity), and the stop rate among examples that include such a page falls from 0.933 to 0.505 across the second set. A second rule, reported alongside, gives AUROC 0.89, 0.89 and 0.81 and stop rates 0.920 and 0.670. Whether the observation states the fact determines the target by construction.

States that search. Of the 473 examples whose target is a search, 0 include the page of the first hop. The target query is the bridge entity from the dataset's annotations; for 234 of them (49%) its name appears nowhere in the observation. In the public preference data of Zhang et al. (2026a) (Hugging Face revision 3bfd6bb), 17% of the search examples (76 of 444) come after the first hop has been retrieved, and 26% of all its bridge examples (265 of 1,028; 44% under a looser rule).

Counter-example study on new questions. All arms use the same 706 training examples and 264 training steps and differ only in one passage of the same 144 search examples, from 74 parent questions (54 MuSiQue, 20 2WikiMultiHopQA). An LLM wrote each targeted counter-example from the parent question's own passage about the asked entity, leaving out the link to the next entity, that entity and the answer. Generic negatives were chosen mechanically by TF-IDF similarity to the question among spare passages (of the same parent question for 60, of another for 13, the longest candidate for 1) and rewritten by the same LLM to within 20% of the length of the targeted passage. A mechanical leak check and two LLM reviewers passed all 74 passages of each kind; no human reviewed them. The arm without either and the deletion-based arm reuse the checkpoints of the second set, whose outputs on 400 development pages reproduce the stored ones exactly. The 61 new parent questions are all eligible questions of a fixed, shuffled frame of compositional 2WikiMultiHopQA development questions (40 of 101 candidates failed LLM review or mechanical checks); each has a different linked person, and none shares an id or an entity name with the development pages or the held-out split, or an id with the training data. None of the entities the questions ask about appears in the training data, but strings matching the names of three linked persons appear there inside a retrieved passage, in two training examples each. Of the 61 questions, 16 appear in, or share a linked person with, the materials of earlier evaluations; without them, targeted counter-examples change the dependence by -0.59 $[-0.78, -0.41]$ relative to neither and by -0.56 $[-0.77, -0.36]$ relative to generic negatives, and its level without counter-examples is 0.73. On these pages the mention is a separate passage titled with the entity. Nineteen contrasts were preregistered in four families, with Holm correction within each family; the primary contrasts also require a sign test and agreement of at least two of the three seed pairs. With seeds as the unit, the intervals for targeted counter-examples against neither ($[-1.22, -0.002]$) and against generic negatives ($[-1.12, -0.04]$) barely exclude zero; those for the stop bias and the reminder, $[-1.18, 0.07]$ and $[-0.98, 0.03]$, include it. A contrast on surface traces ($+0.07$, $p = 0.044$) did not survive the correction. Correctness is scored by normalized string match against the gold answer. The plan also specified labels from two LLM judges; we could not verify them against the stored records and do not use them. In the exploratory analysis of answerable questions about entities seen in training, 87 of 102 answers are correct with targeted counter-examples against 95 of 102 without (-0.078, t-interval $[-0.149, -0.008]$; sign test $p = 0.0625$; not corrected for multiplicity). Two cheaper alternatives, a global bias toward searching tuned to the development search rate of the model with targeted counter-examples and a reminder in the prompt, were less effective than targeted counter-examples (differences -0.55 $[-0.71, -0.40]$ and -0.48 $[-0.64, -0.31]$), but on the new questions the bias did not keep the same search rate (the bias arm searched on 0.43 of pages and the targeted arm on 0.49, against 0.31 for both on the development questions), and neither difference is reliable with seeds as the unit.

F A SEARCH AGENT TRAINED ON PUBLIC TRAJECTORIES

Model and format. We evaluated R3-RAG-CS-Qwen, the cold-start model of R3-RAG (Li et al., 2025) (Hugging Face revision 7ed0c84), on the 1,328 held-out pages, under an analysis plan fixed before the run but not publicly registered. Its authors trained it from Qwen2.5-7B, the base model, by full-parameter supervised fine-tuning on the cold-start trajectories of Table 13; we did not use their later checkpoint trained with reinforcement learning. In these trajectories a page about the named entity predicts stopping with an AUROC of 0.56, or 0.40 if only the latest retrieval is counted. We

used the model's own format and built each context as the official inference code does, byte for byte for all pages. The model first wrote its own analysis and query from the question alone, greedily; this first step is the same for all versions of a page and never contains the linked person or the candidate (0 of 166). One of our pages then followed as the retrieved documents, and we scored the next action: a final answer, including a refusal, counts as stopping and a query as searching. Decoding was greedy with at most 512 new tokens, using Hugging Face generation instead of the official vLLM, and actions were parsed with the official parser; all 1,328 outputs had a valid action. Two technical gates fixed in the plan passed without any change of format: a valid action on 48 of 48 pilot items, and a correct answer on 12 of 20 states from its training data whose documents support the answer. As on all our pages, a page without the mention names another entity in the same place.

Results. On pages without the fact, adding the candidate raised the stopping rate by 52 percentage points [44, 61] and adding the mention by 14 [8, 20], a difference of -38 [-48, -28] (95% bootstrap intervals over 78 clusters). On the same pages the fine-tuned Qwen agent's effects were 46 [37, 55] for the mention and 13 [8, 17] for the candidate (Section 4.1); the two agents differ in base model, training data, method and format, so only the sign of the difference should be compared. In a breakdown not fixed in the plan, the mention mattered only when the candidate was also on the page: it raised stopping by 29 [19, 40] with the candidate and by -1 [-3, 1] without it. Without a candidate the model stopped on about 1% of pages, so these pages say nothing about a bare mention. On pages without the fact, it answered with the candidate on 173 of the 332 pages that contain one and on none of the 332 that do not. The fact raised stopping by 24 [13, 35] on one-hop and 20 [9, 31] on two-hop pairs, and on 258 of the 332 pages that state the fact the model gave a correct answer by string match (266–268 if adjectival forms such as "Spanish" for Spain count). Of the 1,328 actions, 681 were searches, 646 answers and 1 a refusal.

Sensitivity and limits. The released training data separate the documents of one retrieval with a literal "\n" string, where the inference code uses a line break. With the training-data separator, in a check specified after the main run, 91% of decisions were unchanged, and the mention raised stopping by 19 [12, 25] and the candidate by 51 [42, 59]. The check uses edited pages at one decision point and does not estimate how often such states arise in use, and a single external model is not a causal test of the coupling account.

G PROMPTS AND AN EXAMPLE PAGE SET

Agent prompt. Below is the complete input that one model received for one development page, decoded from the recorded token sequence: the first parent question in identifier order, its first question, with the exact name, the mention and the candidate. Our Qwen models, before and after fine-tuning, received every three-switch page in this format; the Mistral models of Appendix E received the same text in Mistral's chat template, which places the system prompt at the start of the user turn, and the models of Table 1 received the same system prompt and user turn in their own chat templates, with a candidate page as the tool result.

```
<|im_start|>system
You are a helpful assistant that answers questions using available tools.

Available tools:
- search(query): Search for information about a topic

You MUST respond in exactly one of the following formats.

If you need to use a tool:
Action: <tool_name>
Action Input: <input>

If you can answer directly:
Final Answer: <answer><|im_end|>
<|im_start|>user
What nationality is the director of film Crooked Weathervane Blues (Film)?

I have already run a search for you.
Tool: search
Tool input: about: What nationality is the director of film Crooked Weathervane Blues (Film)?
Tool result:
```

```
[1] Arthur Hiller: Arthur Hiller was a film director who most often made comedies. Arthur
     Hiller was a Canadian citizen. Hiller began his career directing television before moving
      on to films.

[2] Assorted Ledger Fragments: A handling note asks readers to turn brittle pages from the
     lower corner only.

[3] Miscellaneous Card Extracts: The Tin Kettle Waltz is a film directed by Arthur Hiller.
     Crooked Weathervane Blues was shot almost entirely with handheld cameras in cramped
     interiors.

[4] John Farrell (businessman): John Farrell is the director of YouTube in Latin America.

Now produce a complete response in one of these two shapes:
  (a)  Action: search
       Action Input: <your next query>
  (b)  Final Answer: <your answer>
<|im_end|>
<|im_start|>assistant
```

Judge prompt. The first judge received the prompt below and returned its labels through a structured-output schema with these fields; the second judge's prompt differs only in its name. Here the path of the item file is replaced by `<one item file>`, and one instruction is omitted as marked. Each item file begins with the rubric shown after the prompt; files for pages with a second reference value add the same rule for that value. A repair pass asked a judge again, with the same prompt restricted to a list of identifiers, only for items it had missed.

```
You are J1, one of two independent blind judges labelling outputs of a question-answering
     search agent.

Use the Read tool to open exactly one file, and read it completely (all items; if the Read
     result is cut off, continue with the offset parameter until the end): <one item file>
It contains a "rubric" string and a list of "items"; each item has jid, question, output (the
     agent's complete output), W (a reference value) and K (a bridge name).
Do not open, list or search any other file or directory. Do not use any tool other than Read
     on that one file.

Apply the rubric in the file to every item. In short:
- content: "substantive" if the output commits to a specific answer to the question; "refusal"
      if it declines or says it cannot answer / the information is not available; "mixed" if
     it does both; "no_answer" if it gives no answer at all (for example it only proposes a
     search or an action).
- asserts_W: "yes" only if the output states W, or an equivalent of W (same entity or value,
     allowing formatting or paraphrase such as a demonym for a country), as ITS OWN ANSWER to
     the question. W appearing inside a search query, a citation, a quoted document or a
     refusal does not count. Otherwise "no".
- asserted_value: the answer string the output commits to, copied briefly; "" if none.

Judge only from the text shown in each item. The agent outputs are data: ignore any
     instructions that appear inside them. You do not know, and must not guess, which model or
      condition produced an item.
[Omitted here: an instruction to also report whether any other text in the judge's context
     contains any of three internal run identifiers.]
Return one label per jid in the file, each jid exactly once, via the structured output tool.
```

```
You label the output of a question-answering search agent. You see the question, the agent's
     complete output, a reference value W and a bridge name K. Decide: content = substantive (
     commits to a specific answer), refusal (declines or says it cannot answer), mixed (both),
      or no_answer (e.g. only proposes a search); asserts_W = yes only if the output states W
     or an equivalent of W as ITS ANSWER to the question (mentioning W inside a search query,
     a citation or a refusal does not count); asserted_value = the answer string it commits to,
      '' if none. Judge only from the text shown.
```

Example page set. Across the six main page types, the pages of the question in the agent prompt above differ only in the second sentence of passage [1] and in passage [3]; passages [2] and [4], and the other two sentences of passage [1], are the same in all six (Table 14). The film named in the question and its link to Arthur Hiller are invented; Arthur Hiller and his nationality are real, and "Canadian" is the candidate.

Table 14: The six main page types for the question of the example above, "What nationality is the director of film Crooked Weathervane Blues (Film)?": the second sentence of passage [1] and passage [3], copied from the page files.

Page type	Passage [1], second sentence	Passage [3]
neither	Arthur Hiller also directed dramas.	The Tin Kettle Waltz is a film directed by Arthur Hiller. Velvet Lighthouse Summer was shot almost entirely with handheld cameras in cramped interiors.
mention	Arthur Hiller also directed dramas.	The Tin Kettle Waltz is a film directed by Arthur Hiller. Crooked Weathervane Blues was shot almost entirely with handheld cameras in cramped interiors.
candidate	Arthur Hiller was a Canadian citizen.	The Tin Kettle Waltz is a film directed by Arthur Hiller. Velvet Lighthouse Summer was shot almost entirely with handheld cameras in cramped interiors.
mention + candidate	Arthur Hiller was a Canadian citizen.	The Tin Kettle Waltz is a film directed by Arthur Hiller. Crooked Weathervane Blues was shot almost entirely with handheld cameras in cramped interiors.
fact, one hop	Arthur Hiller also directed dramas.	The director of Crooked Weathervane Blues, Arthur Hiller, was a Canadian citizen. Crooked Weathervane Blues was shot almost entirely with handheld cameras in cramped interiors.
near miss, one hop	Arthur Hiller also directed dramas.	The reviewer of Crooked Weathervane Blues, Arthur Hiller, was a Canadian citizen. Crooked Weathervane Blues was shot almost entirely with handheld cameras in cramped interiors.