TO CONTINUE OR TO COMMIT: EVIDENCE–ANSWER TRACING OF AGENT TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep-search agents address complex search tasks in an iterative search-and-reasoning process until the evidence is judged sufficient and an answer is committed. Evaluating an agent's search-and-reasoning trajectory is important for understanding how it arrives at a final answer, but it is also challenging because the correctness of the outcome depends not only on what the agent does, but also on what information it has acquired and when it acquires it. Existing process signals mostly focus on evaluating the action of an agent, while the acquired information of the agent is largely ignored. This confounds what an action contributes with what the agent already knew. To bridge this gap, we propose *Evidence–Answer Tracing* (EAT), which, to our knowledge, is the first attempt to integrate an agent's action with its information state along the trajectory by borrowing the resource-rational theory. Specifically, two events of a target trajectory are determined by comparing against the ground-truth trajectory: *evidence exposure*, whether the gold documents have surfaced, and *answer surfacing*, whether the gold answer has appeared. This principled framework re-evaluates existing deep-search agents on a fair and unified ground. On 25 state-of-the-art systems on the BrowseComp-Plus, we find that the outcome is largely decided before final commitment. Over 99% of the autonomous-commitment accuracy gap between the strongest and weakest systems lies in which evidence–answer state they reach before committing, and under 1% in within-state conversion, that is, in how well a system answers once that state is reached. The systems that make this progress behave differently along the way: stronger systems redirect when evidence is absent and favor reading over re-searching once partial evidence is available. Once agent and question are held fixed, effort counts reverse sign and LLM-judged behaviors lose much of their association with success, whereas evidence-grounded signals such as the EAT score keep theirs.

1 INTRODUCTION

Consider a deep-search agent that has surfaced a plausible answer but has verified only part of the evidence it needs. It can open one of the documents already in its results, issue a new query, or commit. Two turns later, every gold document is in view and the answer appears in its own reasoning. The agent faces the same three options, but the right one has changed. A deep-search agent makes this choice dozens of times per question, and much current work targets this choice. Reward shaping penalizes searching past the point of sufficiency (Sun et al., 2026; Tang et al., 2026). Controllers decide when to stop (Roh & Han, 2026; Soudani et al., 2026). Training regimes push the number of turns higher (Gao et al., 2025; Lu et al., 2025). None of this work tells us whether a decision taken at a given moment by a deployed system leads to final success. Nor does it tell us where along the trajectory success is decided, at the final step, when the agent reads its evidence and writes an answer, or earlier, when it retrieves that evidence in the first place.

Three lines of work approach this question, and each stops short of it. The first evaluates the process rather than the answer: behavior taxonomies scored by an LLM judge (Jin et al., 2026; Ou et al., 2025), trajectory-level dimensions such as efficiency and grounding (Chen et al., 2026a), and human step labels (Fan et al., 2026). These signals score what the agent does, not what it has found, and their reliability is in doubt: LLM judges of research-agent trajectories fail perturbation tests (Wang et al., 2026), behavior labels follow the label menu (Mazaheri & Mazaheri, 2026), and the correlation between trajectory length and success in coding agents reverses once the task is held fixed (Mehtiyev

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.

& Assunção, 2026). The second line studies the decision itself, from adaptive retrieval (Jiang et al., 2023; Asai et al., 2024; Su et al., 2024) to over- versus under-search (Wu et al., 2025; Xie et al., 2026) and stopping rules (Park et al., 2025; Fang et al., 2026; Roh & Han, 2026; Zhang et al., 2026); training work already uses gold-document coverage (Zheng et al., 2025) and gold-answer availability (Li et al., 2026b; Sun et al., 2026) as rewards. This line shows the signals are informative and how to act on them when the policy can be retrained, but does not measure what they say about the decisions deployed systems actually took. The third line is concurrent: on the same benchmark, Liu et al. (2026) trace per-episode gold-document recall for six agents and list evidence-driven stopping as future work, Soudani et al. (2026) trace per-step recall and answerability for seven agents, and Hamidi Rad et al. (2026) follow gold documents through a surfaced, previewed, and read funnel. None combines evidence with answer surfacing into a joint state, evaluates the action taken from that state, or asks whether systems convert the same state at different rates.

To fill that gap, this paper proposes *Evidence–Answer Tracing* (EAT), in which we evaluate each decision against the information the agent had when it made the decision. EAT records two events along every logged trajectory. *Evidence exposure* records whether none, part, or all of the gold documents have appeared in a result list or an opened page. *Answer surfacing* records whether the normalized gold answer has appeared in the agent's visible context. Both events are computed from the log and the benchmark annotations, are never shown to the agent, and cannot be undone. Together with the number of turns remaining, they place every SEARCH, VISIT, or COMMIT decision in one of six states. This coordinate applies offline to systems whose logged trajectories and benchmark annotations are available. On this coordinate we estimate, for each system, how often each action taken from each state is followed by a correct final answer. We cross-fit the estimates by question so that no trajectory scores itself. All estimates are observational. They describe how actions and outcomes co-occur under matched state, agent, and budget; they do not estimate the effect of replacing one action with another.

We apply EAT to 5,927 trajectories from 25 systems on a stratified subset of BrowseComp-Plus (Chen et al., 2026b). We have the following findings:

- **The outcome is largely decided before final commitment.** The 65.6-point autonomous-commitment accuracy gap between the strongest and weakest tiers decomposes into 77% from evidence exposure, 22% from answer surfacing, and under 1% from the final step. Once all gold documents are exposed and the answer has surfaced, the tiers with enough such commitments to compare (TOP5 and MID) answer correctly at nearly the same rate, 92.9% and 92.0%. Systems differ in the state they reach before committing, not in what they make of it.
- **State progression is associated with redirecting and reading, not more searching.** Stuck without evidence, strong systems change the direction of 71% of their queries and recover in 54% of such episodes; weak systems repeat 48% of their queries and recover in 0.6%. With partial evidence, opening a retrieved document is associated with 2.8 points higher success than committing and issuing another query with 6.3 points lower; once evidence and answer are in hand, a further search is associated with 7.4 points lower success.
- **Evaluation should follow the evidence and answer state, not effort-like behaviors.** When agent and question are held fixed, LLM-judged behavior and step-quality labels retain only 15–57% of their association with success, and effort counts reverse sign. Gold-document recall and the EAT decision value retain their association and carry complementary information. A process signal should therefore be reported under question and agent controls, next to an evidence-grounded baseline.
- **An offline instrument for measuring all of the above from logged trajectories.** EAT provides a deterministic evidence–answer coordinate, a state-conditioned and cross-fitted estimate of decision value, and a progressive audit that places any trajectory-level process signal under the same controls. It applies to systems with the required trajectory logs and benchmark annotations.

Section 2 defines the coordinate and the estimators, Sections 3 and 4 report the cohort and the findings, and Section 5 returns to the three lines of related work.

2 EVIDENCE–ANSWER TRACING

Setup. An agent m answers a question q over a fixed Web corpus. At turn t it observes its history H_t and chooses an external action $a_t \in \{\text{SEARCH}, \text{VISIT}, \text{COMMIT}\}$: SEARCH issues a query and returns ranked results, VISIT opens a surfaced document, and COMMIT submits a final answer and ends the trajectory τ, whose correctness is $C(\tau) \in \{0, 1\}$. The common harness caps every trajectory at $K = 100$ turns, so $R_t = K - t$ is the number of turns remaining under that cap before a_t; K is not shown to the agent, so R_t is a trajectory-position coordinate under the common cap rather than a scaffold's internal budget (Appendix B.1). The benchmark provides a gold answer $y(q)$ and a set of gold documents $\mathcal{G}(q)$ for each question; both are used only for offline analysis and never exposed to the agent (Appendix A).

Figure 1 gives an overview: EAT maps every decision to an evidence–answer state (§2.1), locates capability differences on that state space (§2.2), estimates the association of each action with success under matched state, agent, and budget (§2.3), and audits existing process signals under the same controls (§2.4). The states are deterministic; only the decision values are estimated.

2.1 OBSERVABLE STATE

EAT represents the information available before action a_t as $S_t = (E_t, A_t)$. Let $\mathcal{G}_t(q) \subseteq \mathcal{G}(q)$ be the gold documents exposed to the agent before a_t, through a retrieval result or an opened document, and let $\widetilde{H}_t$ be the trajectory prefix visible at action selection, including the current reasoning step. Then

$$E_t = \begin{cases} E_0, & |\mathcal{G}_t(q)| = 0, \\ E_1, & 0 < |\mathcal{G}_t(q)| < |\mathcal{G}(q)|, \\ E_2, & |\mathcal{G}_t(q)| = |\mathcal{G}(q)|, \end{cases} \qquad A_t = \begin{cases} A_0, & \mathrm{norm}(y(q)) \notin \widetilde{H}_t, \\ A_1, & \mathrm{norm}(y(q)) \in \widetilde{H}_t, \end{cases} \tag{1}$$

where $\mathrm{norm}(\cdot)$ is the benchmark's normalized answer-matching procedure. E_t records no, partial, or complete exposure to the gold evidence, and A_t whether the gold answer has appeared in the visible prefix. Both coordinates are computed from the logged trajectory and the benchmark annotations alone, so the same six states apply to every system whose logs are available. Exposure and surfacing cannot be undone, so both coordinates are non-decreasing along a trajectory. Both are benchmark-relative: E_t records that annotated gold documents have appeared in the agent's result lists or opened pages, not what the agent has read or retained, and A_t records literal surfacing of the gold string. Together with R_t, the pair (S_t, R_t) is the coordinate used in all analyses below.

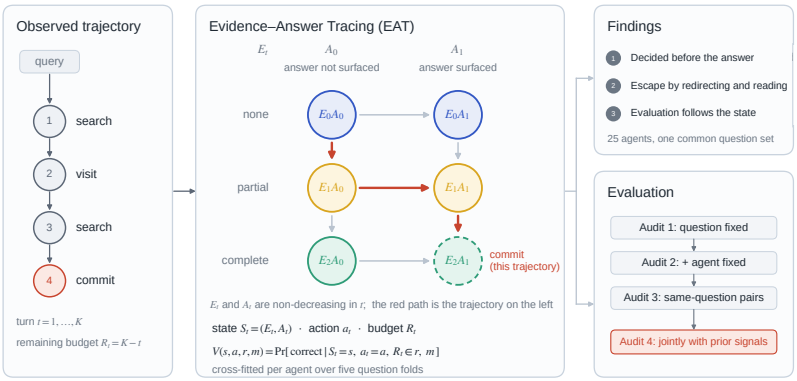

Figure 1: **Overview of EAT.** Each decision in a logged trajectory (left) is mapped deterministically to one of six states defined by gold-document exposure and answer surfacing (middle), with the remaining turns retained as a separate coordinate. The state distribution at commitment locates capability differences (Finding 1), per-system decision values describe which actions lead out of each state (Finding 2), and the same coordinate supports the progressive audit of process signals (Finding 3; right).

2.2 WHERE CAPABILITY DIFFERENCES ARISE

EAT gives a common representation for asking *where* along the task-solving process capability differences arise. We group agents into tiers by overall accuracy and compare them through the terminal-state distribution $\pi_g(s)$, the share of tier g's autonomous commitments made from state s, and the within-state correctness $a_g(s)$, the accuracy of tier g given that it commits from s. Autonomous-commitment accuracy is $\mathrm{Acc}_g = \sum_s \pi_g(s)\, a_g(s)$, and the corresponding gap between tiers g and g' decomposes symmetrically as

$$\mathrm{Acc}_g - \mathrm{Acc}_{g'} = \underbrace{\sum_s \big(\pi_g(s) - \pi_{g'}(s)\big)\, \frac{a_g(s) + a_{g'}(s)}{2}}_{\text{state attainment}} + \underbrace{\sum_s \frac{\pi_g(s) + \pi_{g'}(s)}{2}\, \big(a_g(s) - a_{g'}(s)\big)}_{\text{within-state conversion}}. \tag{2}$$

The first term is the part of the gap explained by reaching different states before commitment; the second is the part explained by converting the same state at different rates. Since the terminal state is the cumulative result of every earlier SEARCH and VISIT, the first term absorbs those decisions and the second isolates the final step. We further factor the gap into the evidence-state distribution $P(E)$, answer surfacing given evidence $P(A \mid E)$, and correctness given both $P(C \mid E, A)$, averaging over all replacement orders (Appendix C). Both decompositions are accounting identities; they describe where the gap sits, not why capability differs.

2.3 STATE-CONDITIONED DECISION ANALYSIS

For an observed decision we define the decision–outcome association

$$V(s, a, r, m) = P(C = 1 \mid S_t = s,\ a_t = a,\ R_t \in r,\ m), \tag{3}$$

where r is an interval of remaining budget: V is the observed success rate of trajectories in which agent m took action a from state s with remaining budget in r. We call V the decision value as shorthand for this observational conditional success rate; it is not a causal action value. To compare actions we hold everything else fixed. For a decision t of trajectory τ, the contrast between a continuation action $a \in \{\text{SEARCH}, \text{VISIT}\}$ and COMMIT is

$$\begin{aligned} \Delta_a(t) &= V(s_t, a, r_t, m) - V(s_t, \text{COMMIT}, r_t, m), \\ \bar{\Delta}_a(s) &= \frac{1}{|\mathcal{T}(s)|} \sum_{\tau \in \mathcal{T}(s)} \frac{1}{|\mathcal{D}_\tau(s)|} \sum_{t \in \mathcal{D}_\tau(s)} \Delta_a(t), \end{aligned} \tag{4}$$

where $\mathcal{D}_\tau(s)$ is the set of decisions of τ taken from s and $\mathcal{T}(s)$ the set of trajectories with at least one such decision. The contrast is averaged first within each trajectory and then across trajectories, so a trajectory with fifty searches from E_0A_0 counts the same as one with two; restricting $\mathcal{D}_\tau(s)$ to one budget interval gives the budget-resolved contrast. V is estimated as a cell mean within agent $\times$ state $\times$ action $\times$ budget-interval cells with five question-level folds, so that decisions from a held-out question never contribute to their own estimates; cells below a minimum support are left unscored rather than extrapolated, and uncertainty comes from bootstrap over trajectories. Both V and the contrasts are observational: they describe how actions and outcomes co-occur under matched conditions, not the effect of replacing one action with another (Appendix A).

2.4 UNIFIED PROCESS-SIGNAL AUDIT

We next ask whether existing process signals remain informative once the confounders that EAT makes explicit are controlled. A signal enters the audit if it can be reconstructed from the logged trajectories and applies to both successful and failed runs (Appendix E). Every signal is projected onto the same agent–question trajectories and passed through four progressively stronger views. For a trajectory-level signal X, with Q denoting question and A agent fixed effects,

$$\text{AUDIT 1}: \ \mathrm{Assoc}(X \mid Q), \qquad \text{AUDIT 2}: \ \mathrm{Assoc}(X \mid Q, A), \tag{5}$$

where $\mathrm{Assoc}(X \mid \mathcal{C})$ is the association of X with final correctness after removing the additive fixed effects in $\mathcal{C}$. Audit 1 asks whether larger signal values remain associated with success among trajectories on the same question; Audit 2 additionally removes persistent agent differences, so that a

signal cannot score well merely because strong agents produce more of it. Audit 3 drops the linear specification and compares every correct–incorrect pair on the same question, reporting how often the signal ranks the correct trajectory higher. Audit 4 enters the prior signal and EAT jointly under the Audit-2 controls and reports what each retains. To place EAT on the same trajectory-level unit, we aggregate its supported out-of-fold decision values,

$$Q_{\mathrm{EAT}}(\tau) = \frac{1}{|\mathcal{T}_\tau|} \sum_{t \in \mathcal{T}_\tau} \widehat{V}(S_t, a_t, R_t, m), \tag{6}$$

where $\mathcal{T}_\tau$ is the set of supported decisions in τ; because $\widehat{V}$ is cross-fitted, the question containing τ never contributes to its own score.

3 Experimental Setup

Cohort. All analyses use one common cohort: 242 questions from a proportional stratified sample of BrowseComp-Plus and 25 systems spanning accuracy .004–.821, giving 5,927 trajectories and 65,301 external decisions. Some runs have incomplete intermediate trajectory logs because of cost or access constraints and are therefore unavailable for process-level analysis. Every system uses the same retriever (dense retrieval over a Qwen3-Embedding-8B index, $k = 5$ documents per SEARCH with 512-token snippets; VISIT returns the full document) and the same 100-turn cap (Appendix B.1). 5,672 trajectories end in an autonomous commitment; the remaining 255 end without a parsable answer, 101 of them at the cap, and are scored as incorrect. Final correctness is a three-grader majority under the official grading template. Systems are grouped into TOP5 (the five highest-accuracy systems, .579–.821), WEAK (below .10), and MID (the remaining ten). Sampling, exclusions, and the system inventory are in Appendix B.

Decision values. The remaining budget R_t is partitioned into five intervals, 1–62, 63–82, 83–92, 93–98, and 99–100 turns, chosen to balance decision counts and then held fixed. V is estimated within system $\times$ state $\times$ action $\times$ interval cells with a minimum training-cell support of 10; 96.18% of eligible decisions receive a supported out-of-fold estimate. Contrasts use 2,000 trajectory-level bootstrap resamples; terminal-state shares and within-state accuracies use question-level bootstrap intervals (Appendix A).

Stuck episodes and query behavior. A trajectory is *stuck* if it spends five or more consecutive non-terminal turns in E_0A_0 and issues at least five queries. Consecutive queries are compared by token Jaccard similarity: a query is a *repeat* of its predecessor at similarity $\geq .85$ and a *pivot* at similarity $\leq .25$. A stuck trajectory *escapes* if it terminates in any state with $E \geq E_1$. These quantities are computed from the logs alone and reported descriptively (Appendix D.5).

Audited signals. Eleven prior signals are reconstructed on the cohort: gold-document recall (the benchmark's Recall metric) and interaction counts from the logs; the four Behavior Priming behaviors and the four AgentDiagnose competencies, re-scored with their official prompts and a common judge (GPT-4.1); and the mean AgentProcessBench step label. All signals are standardized; inference is question-clustered (Audit 1), two-way clustered (Audits 2 and 4), or paired question-level bootstrap (Audit 3). Exclusions are listed in Appendix E.

4 Results

4.1 The Outcome Is Largely Decided Before Final Commitment

Once the same state is reached, tiers are nearly indistinguishable. From E_2A_1, TOP5 and MID commit correctly 92.9% and 92.0% of the time (Table 1); WEAK reaches this state in only six runs, all correct, too few to compare. From E_0A_0, every tier is at or near zero. The one state with a clear same-state difference is E_1A_1, where the answer has surfaced but the evidence is still partial: TOP5 commits correctly in 91.7% of cases and WEAK in 65.2% (Figure 2b).

Strong and weak systems commit from different states. TOP5 commits from E_2A_1 in 53.1% of runs and WEAK in 0.3%; conversely, 92.8% of WEAK commitments are made before any gold document has been exposed, against 17.0% for TOP5 (Figure 2a, Table 1). Weak systems fail to retrieve the evidence and therefore never surface a candidate answer.

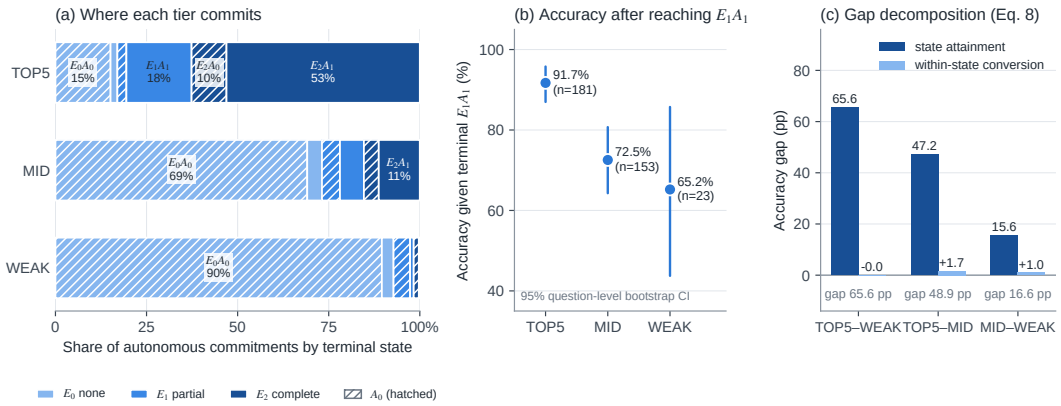

Figure 2: **Capability differences are concentrated in the state reached before commitment.** (a) Terminal-state distribution of autonomous commitments per tier; darker blue marks more complete gold evidence, hatching marks states where the answer has not surfaced. (b) Accuracy after committing from E_1A_1, the one state with a clear same-state difference, with 95% question-level bootstrap intervals. (c) Symmetric decomposition of each tier gap (Eq. 2) into state attainment and within-state conversion.

Table 1: **Where tiers differ.** Share of autonomous commitments made before any gold document is exposed (E_0), with the gold answer surfaced (A_1), and from the complete state (E_2A_1), with 95% question-level bootstrap intervals, and accuracy of commitments from the two answer-surfaced states (counts in parentheses). Full six-state tables are in Appendix C.

Tier	E_0 commit (%)	A_1 commit (%)	E_2A_1 commit (%)	E_1A_1 accuracy	E_2A_1 accuracy
TOP5	17.0 [13.9,20.2]	72.8 [68.6,77.0]	53.1 [48.3,57.8]	.917 (181)	.929 (537)
MID	73.1 [70.4,75.7]	21.9 [19.2,24.4]	11.3 [9.6,13.0]	.725 (153)	.920 (262)
WEAK	92.8 [90.9,94.4]	4.5 [3.2,5.9]	0.3 [0.0,0.6]	.652 (23)	1.000 (6)

The autonomous-commitment accuracy gap is almost entirely attainment. Eq. 2 assigns essentially the whole 65.6-point TOP5–WEAK gap to state attainment (within-state conversion: -0.04 points); the attainment share is 96.6% for TOP5–MID and 93.8% for MID–WEAK (Figure 2c). Separating the two events, evidence exposure is the largest factor in every gap (68–87%), answer surfacing matters most between MID and WEAK (26%), and the final step never exceeds 6% (Table 2). Earlier commitment by weak systems does not explain this (Appendix C). The gap lies in which state a system reaches, not in how it converts it; which actions accompany reaching that state is the next finding.

4.2 State Progression Is Associated with Redirecting and Reading, Not Searching More

Stuck without evidence, strong systems pivot and weak systems repeat. Among the 1,332 trajectories that spend five or more turns in E_0A_0, TOP5 systems change the direction of 70.6% of their queries and recover in 54.4% of episodes, whereas WEAK systems repeat the previous query 48.1% of the time and recover in 0.6% (Appendix D.5). Across systems, pivot rate and recovery rate correlate at $\rho = .82$. This is a between-system contrast: it describes what strong and weak systems do in the same state, not the effect of pivoting within a system.

With partial evidence, reading and re-searching diverge. From E_1A_0 and E_1A_1, VISIT sits above COMMIT ($+2.4$ and $+2.8$ points) while SEARCH sits below (-7.8 and -6.3; Figure 3a). When part of the evidence is already in the results, opening it is associated with success and issuing a new query is not. Effort measures that pool the two actions as "continuation" cannot see this.

Table 2: **Three-factor decomposition of the autonomous-commitment tier gaps.** Contribution of the evidence-state distribution $P(E)$, answer surfacing given evidence $P(A \mid E)$, and correctness given both $P(C \mid E, A)$, averaged over all replacement orders (Appendix C).

Contrast	Total gap	Evidence exposure	Answer surfacing	Final step
TOP5–WEAK	65.6 pp	50.30 pp (76.7%)	14.74 pp (22.5%)	0.53 pp (0.8%)
TOP5–MID	48.9 pp	42.72 pp (87.3%)	4.58 pp (9.4%)	1.62 pp (3.3%)
MID–WEAK	16.6 pp	11.31 pp (67.9%)	4.28 pp (25.7%)	1.05 pp (6.3%)

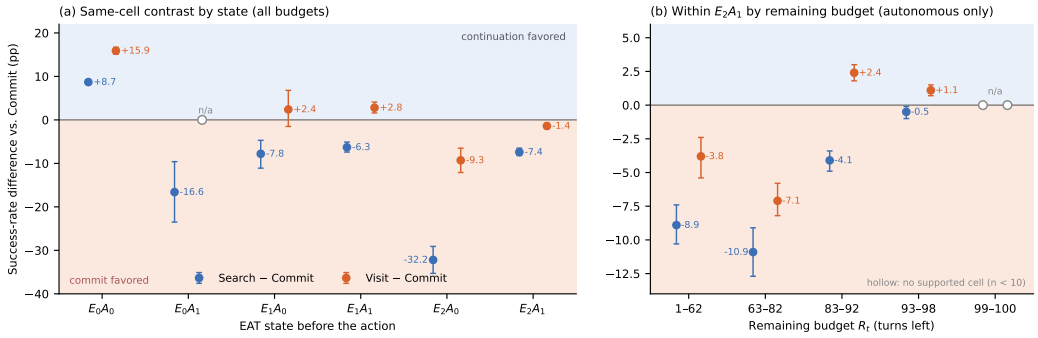

Figure 3: **The outcome association of continuing depends on the state and on the turns remaining.** (a) Success-rate difference of SEARCH and VISIT relative to COMMIT within the same system, state, and budget interval, averaged over trajectories; above zero favors continuing, below zero favors committing. (b) The same contrast within E_2A_1 by remaining budget, restricted to trajectories that end in an autonomous commitment. Error bars are 95% bootstrap intervals; hollow markers mark cells without support. Values are listed in Tables 9 and 10.

Once evidence and answer are in hand, continuing only costs. From E_0A_0, SEARCH runs 8.7 points above COMMIT, since committing with nothing in hand is close to hopeless for every tier; from E_2A_1 it runs 7.4 points below. No fixed rule survives this reversal: "search more" is right at one end of the state space and wrong at the other, and "commit once an answer appears" is wrong at E_1A_1. Within E_2A_1 the cost grows as the budget runs down, from -0.5 points with 93–98 turns left to -10.9 with 63–82 (Figure 3b, autonomous commitments only; Appendix D.2 separates the mechanical cost of hitting the cap). The cost is one of prolonged delay rather than of a single extra step: TOP5 commitments made on entering E_2A_1 are correct 95.4% of the time, unchanged after up to five further actions, and 87.1% after six or more (Table 11). The reversal holds within TOP5 alone and across support thresholds (Appendix D).

4.3 EVALUATION SHOULD FOLLOW THE STATE, NOT THE EFFORT

Holding the agent fixed removes most of the association. With only the question held fixed (Audit 1), every prior signal is positively associated with success, and search-turn count looks like a useful effort measure at $+6.7$ points per standard deviation (Figure 4, horizontal axis). Once stable agent differences are also controlled (Audit 2), search-turn count reverses to -10.9: within a system, the runs that search more are the runs that fail. The behavior, reasoning, and process-judge signals keep 15–57% of their Audit-1 association (Table 17), well below the diagonal of Figure 4. The Audit-1 values were carried by composition: stronger agents both exhibit these behaviors more and succeed more.

Two signals survive, and they are complementary. Gold-document recall is unchanged ($+32.5 \rightarrow +31.9$) and Q_{EAT} strengthens ($+32.1 \rightarrow +38.8$); without a linear model, the two have the highest same-question concordance (.958 and .961, statistically tied). Entered jointly (Audit 4), each keeps 22–23 points while every other signal keeps at most 5.2 next to EAT: what was eventually found and which decisions were taken along the way are different information. Because Q_{EAT} is fitted

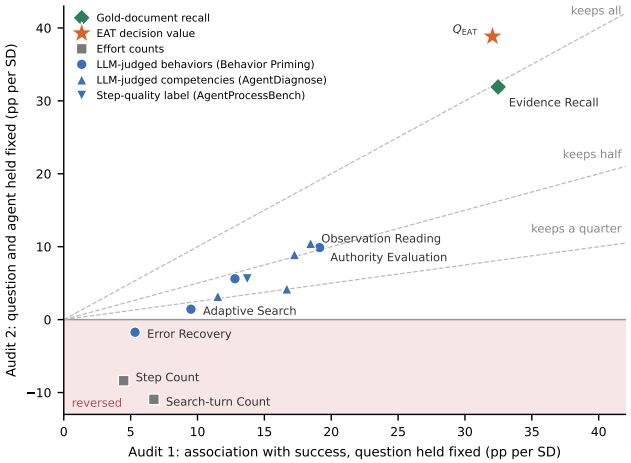

Figure 4: **Evidence-grounded signals survive agent control; judged and effort signals do not.** Each point is a trajectory-level signal reconstructed on the 5,927 trajectories. Its horizontal position is the association with correctness among trajectories on the same question (Audit 1); its vertical position is the same association after also holding the agent fixed (Audit 2). Points on the diagonal keep their discrimination. Values are in Table 17.

Table 3: **Progressive audit of representative process signals.** Audit 1 controls question variation; Audit 2 also controls stable agent differences; A2/A1 is their ratio; Audit 3 is same-question correct–incorrect concordance; Audit 4 reports what the prior signal and Q_{EAT} each retain when entered jointly. Associations are percentage points per standard deviation of the signal. All eleven signals are in Table 17 and plotted in Figure 4.

Family	Signal	Audit 1	Audit 2	A2/A1	Audit 3	Audit 4	
					C_Q	Prior	EAT
Evidence	Gold-document recall	+32.5	+31.9	.98	.958	+22.2	+23.1
Effort	Search-turn count	+6.7	−10.9	−1.62	.812	+0.7	+39.2
Behavior	Authority Evaluation	+19.1	+9.9	.52	.790	+4.8	+36.9
Reasoning	Observation Reading	+18.5	+10.5	.57	.829	+5.2	+37.0
Process judge	APB mean step label	+13.7	+5.6	.41	.747	+2.3	+38.3
EAT	Q_{EAT}	+32.1	**+38.8**	**1.21**	**.961**	–	–

on outcome labels while recall is not, we also compare all signals under matched supervision: recall predicts better on concordance, Q_{EAT} on Brier error, and their combination on both, so the complementary result stands (Appendix E.3). A process signal should therefore be reported under question and agent controls and next to an evidence-grounded baseline; otherwise its discrimination cannot be distinguished from knowing which agent produced the trajectory.

5 Related Work

Scoring the process. Process evaluation of search agents began with behavior taxonomies detected by an LLM judge (Jin et al., 2026; Ou et al., 2025), moved to trajectory-level dimensions (Chen et al., 2026a), and replaced the judge with human step labels (Fan et al., 2026); process reward models learn step scores from outcomes (Xi et al., 2025; Li et al., 2026a), and failure-localization work asks where a trajectory went wrong (Zhao et al., 2026; Zhan et al., 2026). All of these score what the agent does; none records what it has. Their reliability is in question: LLM judges of research-agent trajectories fail perturbation tests (Wang et al., 2026), behavior labels follow the label menu (Mazaheri & Mazaheri, 2026), task difficulty confounds behavior–outcome associations (He et al., 2026), and in coding agents the pooled correlation between trajectory length and success reverses within tasks (Mehtiyev & Assunção, 2026). Our audit is the deep-search counterpart of these findings.

When to search and when to stop. Adaptive retrieval asked when a single retrieval is worth making (Jiang et al., 2023; Asai et al., 2024; Jeong et al., 2024; Su et al., 2024); for iterative agents the question became over- versus under-search (Wu et al., 2025; Xie et al., 2026; Tang et al., 2026) and a family of stopping rules: value-based (Park et al., 2025), budget-aware (Fang et al., 2026), coverage-keyed (Roh & Han, 2026), and counterfactual (Zhang et al., 2026). Training work uses the two events we trace as rewards: StepSearch rewards per-step gains in gold-document coverage (Zheng et al., 2025), SE-Search rewards the gold answer entering memory (Li et al., 2026b), and AutoSearch rewards stopping at the earliest step whose intermediate answer matches gold (Sun et al., 2026). We use the same signals as a measurement coordinate on logged trajectories of 25 systems that cannot be re-invoked, and evaluate the decision conditionally on state and budget rather than optimizing it. The two heuristics we test, "search more" (Gao et al., 2025; Lu et al., 2025; Zeng et al., 2026) and "commit once the answer appears" (Sun et al., 2026; Wu et al., 2025), each anchor a published line; we find each right in one state and wrong in another.

Concurrent trajectory diagnostics on BrowseComp-Plus. Liu et al. (2026) compute cumulative gold-document recall per search episode for six agents and show under question fixed effects that recall dominates every behavioral predictor, listing evidence-driven stopping as future work; Soudani et al. (2026) compute per-step recall and, by branching a candidate answer, per-step answerability for seven agents; Hamidi Rad et al. (2026) trace gold documents through a surfaced, previewed, and read funnel. We reach the same evidence coordinate independently and reproduce its dominance (our question-controlled recall effect is within three points of theirs). What we add is the answer coordinate as a second event, the joint state as the unit of analysis, the value of each action taken from each state under a fixed budget, and the finding that conversion of a state is nearly invariant across tiers. Resource rationality (Lieder & Griffiths, 2020; Ding et al., 2026) motivates our conditioning on the budget remaining; we do not fit a cost objective.

6 Limitations and Conclusion

Limitations. All estimates are observational: they describe how actions and outcomes co-occur under matched state, agent, and budget, not the effect of replacing one action with another. Answer surfacing requires the full normalized gold string and therefore misses paraphrases: it detects 78.8% of graded-correct commitments; a variant-aware normalization raises this to 97.2% (Appendix E.4). Exposure counts a document once it appears in a result list, whether or not the agent opens it. The cohort is a 242-question subset of one benchmark; behavior labels are reconstructed with a common judge; and agent fixed effects are additive, so an agent-by-question interaction remains in the residual.

Conclusion. Tracing evidence exposure and answer surfacing across 25 systems shows that capability differences are concentrated before final commitment: the gap between strong and weak systems lies in which evidence–answer state they reach before committing, not in how they convert it. Reaching that state goes with redirecting a stalled search and reading what has already been retrieved, whereas searching more goes with staying stuck; and process signals that ignore this state lose their association with success once the agent is held fixed, while evidence-grounded signals keep theirs. Deep-search agents should therefore be evaluated, and improved, on the evidence–answer state they reach and on what moves them into it, rather than on how much they search.

Reproducibility Statement

All analyses use one fixed cohort: the 242 BrowseComp-Plus question identifiers, the eight excluded identifiers, and the 25-system inventory are listed in Appendix B, together with the harness, retriever, and turn-cap settings under which every trajectory was produced. The EAT state of every decision is a deterministic function of the logged trajectory and the benchmark annotations (Appendix A), and all estimators use fixed question-level folds (seed 2027), a fixed minimum cell support, and fixed budget intervals, whose selection is described in Appendix B.2. Every number in the main text is listed with its interval in Appendices C–E. The analysis code, the per-decision EAT states, and the reconstructed process signals will be released with the paper.

AI USE STATEMENT

Large language models are part of the experimental setup: GPT-4.1 serves as the common judge that rescores the Behavior Priming, AgentDiagnose, and AgentProcessBench signals (Appendix E), Qwen3-32B, DeepSeek-V3.1, and GPT-4o serve as the three graders of final correctness (Appendix B.2), and the 25 evaluated systems are themselves LLM-based agents. In preparing the manuscript, LLM tools were used for retrieval and discovery of related work and to aid and polish the writing. All analyses, numbers, and claims were produced and verified by the authors, who take full responsibility for the content.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511.

Yanyu Chen, Jiyue Jiang, Jiahong Liu, Yifei Zhang, Xiao Guo, and Irwin King. Trace: Trajectory-aware comprehensive evaluation for deep research agents. In *Proceedings of the ACM Web Conference 2026*, pp. 2524–2534, April 2026a. doi: 10.1145/3774904.3792738. URL https://doi.org/10.1145/3774904.3792738.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. Browsecomp-plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22349–22370. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.acl-long.1023. URL https://aclanthology.org/2026.acl-long.1023/.

Wenxuan Ding, Nicholas Tomlin, and Greg Durrett. Calibrate-then-act: Cost-aware exploration in LLM agents. *arXiv preprint arXiv:2602.16699*, 2026. URL https://arxiv.org/abs/2602.16699.

Shengda Fan, Xuyan Ye, Yupeng Huo, Zhi-Yuan Chen, Yiju Guo, Shenzhi Yang, Wenkai Yang, Shuqi Ye, Jingwen Chen, Haotian Chen, Xin Cong, and Yankai Lin. AgentProcessBench: Diagnosing step-level process quality in tool-using agents. *arXiv preprint arXiv:2603.14465*, 2026. URL https://arxiv.org/abs/2603.14465.

Zhengru Fang, Senkang Forest Hu, Zhonghao Chang, Yu Guo, Yihang Tao, Hongyao Liu, Mengzhe Ruan, Jun Huang, and Yuguang Fang. Inference-time budget control for LLM search agents. *arXiv preprint arXiv:2605.05701*, 2026. URL https://arxiv.org/abs/2605.05701.

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

Radin Hamidi Rad, Amin Bigdeli, Negar Arabzadeh, Sajad Ebrahimi, Charles L. A. Clarke, Benjamin C. M. Fung, and Ebrahim Bagheri. Question's gambit: The first move matters in agentic deep search. *arXiv preprint arXiv:2609.14412*, 2026. URL https://arxiv.org/abs/2609.14412.

Jiawei He, Mengyu Shi, Jie Jia, Xikai Yang, and Dong Sun. What process evaluation of coding agents actually measures: Action, task, and step are three different levels. *arXiv preprint arXiv:2608.22960*, 2026. URL https://arxiv.org/abs/2608.22960.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

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

Jiahe Jin, Abhijay Sai Paladugu, and Chenyan Xiong. Beneficial reasoning behaviors in agentic search and effective training methods to obtain them. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28080–28097. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.findings-acl.1400. URL https://aclanthology.org/2026.findings-acl.1400/.

Dawei Li, Yuguang Yao, Zhen Tan, Huan Liu, and Ruocheng Guo. Toolprmbench: Evaluating and advancing process reward models for tool-using agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12378–12391, July 2026a. doi: 10.18653/v1/2026.findings-acl.602. URL https://aclanthology.org/2026.findings-acl.602/.

Jian Li, Yizhang Jin, Dongqi Liu, Hang Ding, Jiafu Wu, Dongsheng Chen, Yunhang Shen, Yulei Qin, Ying Tai, Chengjie Wang, Xiaotong Yuan, and Yabiao Wang. SE-Search: Self-evolving search agent via memory and dense reward. *arXiv preprint arXiv:2603.03293*, 2026b. URL https://arxiv.org/abs/2603.03293.

Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020. doi: 10.1017/S0140525X1900061X. URL https://doi.org/10.1017/S0140525X1900061X.

Qi Liu, Jiaxin Mao, Fengbin Zhu, and Tat-Seng Chua. Diagnosing search behavior and failure modes in long-horizon search agents. *arXiv preprint arXiv:2608.01913*, 2026. URL https://arxiv.org/abs/2608.01913.

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. DeepDive: Advancing deep search agents with knowledge graphs and multi-turn RL. *arXiv preprint arXiv:2509.10446*, 2025. URL https://arxiv.org/abs/2509.10446.

Parsa Mazaheri and Kasra Mazaheri. AgentAtlas: Beyond outcome leaderboards for LLM agents. *arXiv preprint arXiv:2605.20530*, 2026. URL https://arxiv.org/abs/2605.20530.

Tural Mehtiyev and Wesley K. G. Assunção. Beyond resolution rates: Behavioral drivers of coding agent success and failure. *arXiv preprint arXiv:2604.02547*, 2026. URL https://arxiv.org/abs/2604.02547.

Tianyue Ou, Wanyao Guo, Apurva Gandhi, Graham Neubig, and Xiang Yue. AgentDiagnose: An open toolkit for diagnosing LLM agent trajectories. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 207–215, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-demos.15. URL https://aclanthology.org/2025.emnlp-demos.15/.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-RAG: Value-based retrieval control for iterative RAG. *arXiv preprint arXiv:2510.14337*, 2025. URL https://arxiv.org/abs/2510.14337.

Daeyoung Roh and Donghee Han. HALT: Verification-aware stopping for retrieval-augmented search agents. *arXiv preprint arXiv:2608.02009*, 2026. URL https://arxiv.org/abs/2608.02009.

Heydar Soudani, Elizabeth Lingg, Faegheh Hasibi, and Navid Rekabsaz. When deep research agents stagnate: Enhancing reasoning with retrieval-aware agent control. *arXiv preprint arXiv:2608.15191*, 2026. URL https://arxiv.org/abs/2608.15191.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 12991–13013, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.702. URL https://aclanthology.org/2024.acl-long.702/.

Jingbo Sun, Wenyue Chong, Songjun Tu, Qichao Zhang, Yaocheng Zhang, Jiajun Chai, Xiaohan Wang, Wei Lin, Guojun Yin, and Dongbin Zhao. AutoSearch: Adaptive search depth for efficient agentic RAG via reinforcement learning. *arXiv preprint arXiv:2604.17337*, 2026. URL https://arxiv.org/abs/2604.17337.

Yunbo Tang, Chengyi Yang, Shiyu Liu, Zhishang Xiang, Zerui Chen, Qinggang Zhang, and Jinsong Su. SAAS: Self-aware reinforcement learning for over-search mitigation in agentic search. *arXiv preprint arXiv:2605.29796*, 2026. URL https://arxiv.org/abs/2605.29796.

Leyao Wang, Yanan He, Peng Chen, Asaf Yehudai, Yixin Liu, Rex Ying, Michal Shmueli-Scheuer, and Arman Cohan. Time to REFLECT: Can we trust LLM judges for evidence-based research agents? *arXiv preprint arXiv:2605.19196*, 2026. URL https://arxiv.org/abs/2605.19196.

Peilin Wu, Mian Zhang, Xinlu Zhang, Xinya Du, and Zhiyu Zoey Chen. Search wisely: Mitigating sub-optimal agentic searches by reducing uncertainty. *arXiv preprint arXiv:2505.17281*, 2025. URL https://arxiv.org/abs/2505.17281.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Yajie Yang, Wenxiang Chen, Zhihao Zhang, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Agentprm: Process reward models for llm agents via step-wise promise and progress. *arXiv preprint arXiv:2511.08325*, November 2025. URL https://arxiv.org/abs/2511.08325.

Roy Xie, Deepak Gopinath, David Qiu, Dong Lin, Haitian Sun, Saloni Potdar, and Bhuwan Dhingra. Over-searching in search-augmented large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7714–7739, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.361. URL https://aclanthology.org/2026.eacl-long.361/.

Weihao Zeng, Keqing He, Chuqiao Kuang, Xiaoguang Li, and Junxian He. Pushing test-time scaling limits of deep search with asymmetric verification. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.06135.

Yuhao Zhan, Tianyu Fan, Linxuan Huang, Zirui Guo, and Chao Huang. Why your deep research agent fails? on hallucination evaluation in full research trajectory. *arXiv preprint arXiv:2601.22984*, January 2026. URL https://arxiv.org/abs/2601.22984.

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*, pp. 2049–2059. Association for Computing Machinery, April 2026. doi: 10.1145/3774904.3792235. URL https://doi.org/10.1145/3774904.3792235.

Xiangxin Zhao, Han Li, Shuaiting Li, Tianyi Zhao, Earl T. Barr, Federica Sarro, and He Ye. Failure as a process: An anatomy of cli coding agent trajectories. *arXiv preprint arXiv:2607.09510*, July 2026. URL https://arxiv.org/abs/2607.09510.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. StepSearch: Igniting LLMs search ability via step-wise proximal policy optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21805–21830, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1106. URL https://aclanthology.org/2025.emnlp-main.1106/.

A Formal Setup and Estimation Details

Interaction history. For question q, let $H_t = (q, z_0, a_0, o_0, \ldots, z_{t-1}, a_{t-1}, o_{t-1})$, where z_i is the agent's reasoning, a_i the external action, and o_i the environment observation. At turn t the agent produces z_t and then chooses a_t. The prefix available at action selection is $\widetilde{H}_t = (H_t, z_t)$; because it includes the current reasoning step, A_t records whether the answer is visible at the moment the external action is chosen. A complete trajectory terminates with $\textsc{Commit}(\hat{y})$ or, if the 100-turn cap is reached, without an answer.

Exact EAT state. $\mathcal{G}_t(q)$ is the subset of $\mathcal{G}(q)$ whose document identifiers appear in any retrieval result or opened document within $\widetilde{H}_t$. Answer surfacing requires the full normalized gold-answer string under the benchmark's normalization; answer components are not matched separately. Both coordinates are computed offline and never shown to the agent.

Resource rationality. Resource-rational analysis evaluates a strategy h by expected utility minus expected resource cost, $h^* = \arg\max_h \{\mathbb{E}[u(\text{result})] - \mathbb{E}[\text{cost}(h)]\}$ (Lieder & Griffiths, 2020). We instantiate utility with final correctness and cost with interaction turns, and use the principle only to motivate conditioning on the state and on R_t: if a continuation is not associated with higher success, its turn cost cannot make it preferable for any non-negative cost coefficient. Finding 2 is read as consistent with this criterion, not as a test of the full cost model.

Resource-agnostic variant. To test whether the remaining opportunity carries information beyond the state, we also fit the budget-agnostic variant

$$V^{-R}(s, a, m) = P(C = 1 \mid S_t = s,\ a_t = a,\ m) \tag{7}$$

and compare the held-out predictive error of Eqs. 3 and 7 on the same decisions (Appendix D.3).

Causal scope. Both V and V^{-R} are observational. They describe how actions and outcomes co-occur in logged trajectories under matched conditions; they do not estimate the causal effect of replacing one action with another, which would require re-running the policy. Under the fixed cap, R_t is a conditioning variable, not a randomized treatment.

Cross-fitting and support. All state-conditioned estimators use five folds at the question level, so every decision and trajectory from the same question stays in one fold; the fold assignment is fixed (seed 2027) within each reported comparison. Cells with fewer than 10 training decisions are unsupported and left unscored. Same-cell contrasts average $\widehat{V}(s, \textsc{Search}, r, m) - \widehat{V}(s, \textsc{Commit}, r, m)$ over the supported decisions of a trajectory, then over trajectories; equal-system variants average over trajectories within a system, then over systems.

Out-of-fold notation. Let $\widehat{V}^{(-f)}$ denote the cell-mean estimate fitted on the four folds other than the fold $f(q)$ of question q; Eq. 4 is evaluated with $\widehat{V}^{(-f)}$ and only where both cells are supported. Budget-interval contrasts (Figure 3b) restrict $\mathcal{D}_\tau(s)$ to one interval; equal-system contrasts average Eq. 4 within each system first. Intervals are 95% percentile bootstrap over trajectories ($B = 2{,}000$), or over systems for the within-system check.

B Reproducibility and Analysis Scope

B.1 Harness

All systems run against the same corpus and retriever: dense retrieval over a Qwen3-Embedding-8B index of the BrowseComp-Plus corpus, $k = 5$ documents per SEARCH, each result truncated to a 512-token snippet under the Qwen2.5-7B tokenizer; VISIT returns the full document text capped at 12,000 characters. The unified harness allows $K = 100$ iteration loops, does not expose the cap to the agent, and does not force an answer when the cap is reached. Six systems (the *Official* entries in Table 5) run under their authors' own scaffold and prompts with the same retriever and k; their scaffolds stop earlier (WebShaper at most 40 turns, WebDancer 14–17, R1-Searcher 11). R_t is

defined against $K = 100$ for every system; these six systems commit almost exclusively from E_0 and do not enter the E_2A_1 budget-resolved contrasts.

Name suffixes: MT10K raises the per-turn generation budget from 4,000 to 10,000 tokens; 80K extends the serving window from 40,960 to 81,920 tokens (YaRN factor 4) with a 10,000-token generation budget; K5 sets the official scaffold to $k = 5$, matching the unified harness; Fix3 is the third corrected run of the WebExplorer-8B client.

Of the 255 trajectories without an autonomous commitment, 101 reached the 100-turn cap (99 with a pending tool call, 2 at the LLM-call limit) and 154 ended earlier without a parsable answer: 74 stopped normally but yielded no extractable final answer, 39 were rejected by the serving backend because the request exceeded the context window, 22 were forced to answer at the token limit and failed the answer format, 13 hit the per-turn generation limit, and 6 hit the LLM-call limit. By tier: TOP5 63 at the cap and 16 earlier, MID 36 and 53, WEAK 2 and 85. Two trajectories terminated by an external `APIConnectionError` are removed. All 255 are scored as incorrect.

B.2 QUESTION SAMPLING AND COHORT

The 250 questions are a proportional stratified sample of the 830-question BrowseComp-Plus set; stratification uses only intrinsic question properties, constraint-chain length and answer type (Table 4). The remaining 580 questions are not used in any analysis of this paper. Eight questions on which the three graders disagreed about answer equivalence for identical final-answer strings are removed (query ids 170, 253, 268, 556, 804, 1079, 1121, 1221), leaving 242. All decisions from a trajectory share one final correctness label from the three-grader majority (Qwen3-32B, DeepSeek-V3.1, GPT-4o) under the official grading template. The five budget intervals and the minimum cell support of 10 were chosen on the 26-system, 242-question analysis cohort (before GPT-5.5 was excluded) to balance decision counts across intervals, and were then held fixed for every analysis reported here; the stuck threshold of five turns and the Jaccard cut-offs were likewise fixed once on the same cohort. None of these choices used held-out questions; Table 13 reports the sensitivity of the endpoint contrasts to the support threshold.

Table 4: Marginal distributions (%) of the full 830-question set and the 250-question stratified sample.

Question property	Full 830	Sample 250
Constraint length: L1	36.3	36.0
Constraint length: L2	18.7	19.2
Constraint length: L3	23.4	23.2
Constraint length: L4	21.7	21.6
Answer type: short name	71.9	72.0
Answer type: long phrase	13.9	13.6
Answer type: date	5.1	5.2
Answer type: numeric	4.8	4.8
Answer type: multi-part	4.3	4.4

B.3 SYSTEM INVENTORY AND TIERS

Table 5 lists the 25 systems with accuracy on the 242-question cohort and tier membership. GPT-5.5 is part of the broader evaluation archive but is excluded from every analysis because its logged interface exposes insufficient visible reasoning text for the reproduced process evaluators.

C FINDING 1: FULL CAPABILITY-GAP ANALYSIS AND ROBUSTNESS

Of 5,927 trajectories, 5,672 end in an autonomous commitment; tier membership follows Table 5.

C.1 TERMINAL-STATE DISTRIBUTION, WITHIN-STATE CORRECTNESS, AND DECOMPOSITION

Table 6 gives, for each tier and terminal state, the share of autonomous commitments made from that state ($\pi_g(s)$) and the accuracy of those commitments ($a_g(s)$); these are the inputs to Eq. 2, whose output is Table 7. The six WEAK E_2A_1 commitments are too sparse to support a comparative claim. Changing the reference weights in Eq. 2 from the symmetric average to either one-sided tier

Table 5: Final 25-system analysis inventory and capability tiers on the 242-question cohort.

System	Acc.	Tier	System	Acc.	Tier
Claude Opus 5	.821	TOP5	Qwen3-30B-A3B-Instruct-2507	.116	MID
Kimi-K2.7-Code	.694	TOP5	WebDancer-32B-Official-K5	.112	MID
Claude Sonnet 4.6	.665	TOP5	Qwen2.5-32B-Instruct	.091	WEAK
Kimi-K3	.631	TOP5	WebShaper-32B-Official	.087	WEAK
GLM-5.1	.579	TOP5	Qwen2.5-7B-MT10K	.079	WEAK
GLM-5.2	.559	MID	QwQ-32B	.079	WEAK
DeepSeek-V4-Flash	.446	MID	Qwen3-32B	.079	WEAK
WebExplorer-8B-Fix3	.297	MID	Qwen3-32B-MT10K	.070	WEAK
DeepSeek-V3.2	.215	MID	Qwen2.5-7B	.066	WEAK
WebSailor-32B	.153	MID	WebShaper-32B-Official-MT10K	.062	WEAK
QwQ-32B-MT10K	.136	MID	R1-Searcher-7B-Official	.050	WEAK
WebDancer-32B-Official-MT10K-K5	.124	MID	OffSeeker-8B-DPO-Official80K	.004	WEAK
DeepSeek-R1	.120	MID			

distribution does not alter the conclusion; shares slightly above 100% arise when the within-state component has the opposite sign.

Table 6: Terminal EAT-state distribution among autonomous commitments (share, %) and final correctness conditional on the terminal state (accuracy, with counts). n = 1012 (TOP5), 2327 (MID), 2333 (WEAK).

	TOP5		MID		WEAK	
State	share	accuracy (n)	share	accuracy (n)	share	accuracy (n)
E_0A_0	15.1	.000 (153)	69.1	.034 (1608)	89.6	.013 (2090)
E_0A_1	1.9	.263 (19)	4.0	.553 (94)	3.2	.520 (75)
E_1A_0	2.5	.600 (25)	4.9	.461 (115)	4.5	.566 (106)
E_1A_1	17.9	.917 (181)	6.6	.726 (153)	1.0	.652 (23)
E_2A_0	9.6	.495 (97)	4.1	.379 (95)	1.4	.394 (33)
E_2A_1	53.1	.929 (537)	11.3	.920 (262)	0.3	1.000 (6)

Table 7: Autonomous-commitment accuracy-gap decomposition (Eq. 2) between tiers. One-sided weightings give attainment shares of 97.9–102.2% (TOP5–WEAK), 91.2–102.0% (TOP5–MID), and 90.9–96.6% (MID–WEAK).

Contrast	Total gap	Attainment	Within-state	Attainment share
TOP5–WEAK	65.6 pp	65.6 pp	-0.04 pp	$\approx 100\%$
TOP5–MID	48.9 pp	47.2 pp	$+1.68$ pp	96.6%
MID–WEAK	16.6 pp	15.6 pp	$+1.04$ pp	93.8%

C.2 THREE-FACTOR DECOMPOSITION

To separate the two events, each tier gap is factored into the evidence-state distribution $P_g(E)$, answer surfacing given evidence $P_g(A \mid E)$, and within-state correctness $P_g(C \mid E, A)$, all taken from the terminal-state table (cells without commitments count as accuracy 0), with $\mathrm{Acc}_g = \sum_{E,A} P_g(E)P_g(A \mid E)P_g(C \mid E, A)$. Replacing the three factors one at a time in all 3! orders and averaging the marginal changes (a Shapley decomposition) gives Table 2: TOP5–WEAK 50.30 / 14.74 / 0.53 pp, TOP5–MID 42.72 / 4.58 / 1.62 pp, MID–WEAK 11.31 / 4.28 / 1.05 pp. In every tier gap the evidence-state distribution is the largest factor.

C.3 COMMITMENT TIMING

A possible alternative explanation is that WEAK systems commit earlier and therefore have less time to reach E_2A_1. Let $p_g(b)$ be the commit-turn distribution of tier g and $h_g(b) = P_g(E_2A_1 \mid b)$. The TOP5–WEAK E_2A_1 gap decomposes, as in Eq. 2, into a commit-timing component and a same-time-bin attainment component (Table 8). Across four binnings, attainment within comparable commit-time regions explains 80.2–98.7% of the gap.

Table 8: TOP5–WEAK E_2A_1 attainment-gap decomposition under alternative commit-turn binnings.

Scheme	Bins	Gap	Timing	Same-time attain.	Same-time share
Primary	{1, 2–5, 6–10, 11–20, 21–50, 51–100}	52.81	3.29	49.51	93.8%
Fine early	{1, 2, 3, 4, 5, 6–10, 11–20, 21–50, 51–100}	52.81	10.46	42.34	80.2%
First five together	{1–5, 6–10, 11–20, 21–50, 51–100}	52.81	0.71	52.09	98.7%
Dense early	{1, 2–3, 4–5, 6–10, 11–15, 16–20, 21–30, 31–50, 51–100}	52.81	5.43	47.38	89.7%

D FINDING 2: FULL STATE-CONDITIONED RESULTS AND ROBUSTNESS

D.1 SAME-CELL DIFFERENCES BY STATE AND BY BUDGET

Table 9 lists the values plotted in Figure 3: the six-state contrasts pooled over budgets and the E_2A_1 contrasts by remaining-budget interval. E_0A_1 has no supported VISIT contrast and the 99–100 interval has no supported contrast at all.

Table 9: Same-cell action differences with 95% bootstrap intervals. Positive values favor the continuation action relative to COMMIT. Top: by state, all budgets (Figure 3a). Bottom: within E_2A_1, by remaining-budget interval (Figure 3b).

	Search–Commit	Visit–Commit
E_0A_0	+.087 [+.082,+.092]	+.159 [+.151,+.167]
E_0A_1	−.166 [−.235,−.096]	–
E_1A_0	−.078 [−.111,−.047]	+.024 [−.015,+.068]
E_1A_1	−.063 [−.074,−.051]	+.028 [+.016,+.041]
E_2A_0	−.322 [−.353,−.291]	−.093 [−.121,−.065]
E_2A_1	−.074 [−.082,−.065]	−.014 [−.020,−.008]
E_2A_1, R_t 1–62	−.392 [−.416,−.364]	−.189 [−.214,−.162]
E_2A_1, R_t 63–82	−.164 [−.183,−.147]	−.110 [−.123,−.097]
E_2A_1, R_t 83–92	−.054 [−.062,−.046]	+.018 [+.012,+.023]
E_2A_1, R_t 93–98	−.012 [−.017,−.008]	+.006 [+.003,+.010]
E_2A_1, R_t 99–100	–	–

D.2 THE CAP AND DELAYED COMMITMENT

The cap. The bottom block of Table 9 is computed on all trajectories, including the 255 that end without an answer. Because the cap is hard and answers are not forced, a late SEARCH in such a trajectory is followed by failure mechanically. Table 10 recomputes the E_2A_1 budget-interval contrasts on the 5,672 trajectories that end in an autonomous commitment (training and evaluation both restricted; folds and support rule unchanged) and reports the share of SEARCH decisions in each cell supplied by non-autonomous trajectories. Most of the -39.2 at 1–62 is mechanical; the ordering by remaining budget survives at about a fifth of the size, and the main text and Figure 3b report the restricted values.

Table 10: E_2A_1 contrasts by remaining budget, all trajectories vs. autonomous commitments only, and the share of the SEARCH cell supplied by non-autonomous trajectories (decisions / trajectories).

R_t	S–C, all	S–C, autonomous	V–C, all	V–C, autonomous	non-autonomous share
1–62	−39.2 [−41.6, −36.4]	−8.9 [−10.3, −7.4]	−18.9 [−21.5, −16.3]	−3.8 [−5.4, −2.4]	39.3% / 14.7%
63–82	−16.4 [−18.3, −14.5]	−10.9 [−12.7, −9.1]	−11.0 [−12.4, −9.6]	−7.1 [−8.2, −5.8]	8.6% / 3.6%
83–92	−5.4 [−6.2, −4.6]	−4.1 [−4.9, −3.4]	+1.8 [+1.1, +2.4]	+2.4 [+1.8, +3.0]	2.2% / 1.1%
93–98	−1.2 [−1.7, −0.8]	−0.5 [−1.0, −0.1]	+0.6 [+0.3, +1.0]	+1.1 [+0.7, +1.5]	1.2% / 0.9%

Immediate vs. delayed commitment. Among trajectories that end with an autonomous commitment from E_2A_1, let k be the number of SEARCH or VISIT actions between entering the state and committing. Table 11 reports accuracy by k. Accuracy is flat for one to five further actions and falls only after six or more; WEAK contributes six such commitments, all correct.

Table 11: Accuracy of autonomous commitments from E_2A_1 by the number k of further actions taken after entering the state (n in parentheses).

Tier	$k = 0$	$k = 1$–2	$k = 3$–5	$k \geq 6$	all $k \geq 1$
TOP5	95.4% (87)	96.2% (158)	95.0% (121)	87.1% (171)	92.4% (450)
MID	88.9% (36)	94.3% (87)	94.5% (73)	87.9% (66)	92.5% (226)

D.3 Remaining-Position Ablation

The resource-conditioned estimator (Eq. 3) and the resource-agnostic one (Eq. 7) use the same question-level folds and minimum cell support, and are compared only on decisions supported under both. Restricting to autonomous-commit trajectories yields a common held-out pool of 50,878 decisions from 5,639 trajectories. Errors are averaged within trajectory, then within question, then across questions; paired differences use 10,000 question-level bootstrap resamples (Table 12). Conditioning on R_t reduces question-balanced Brier error from .1130 to .1003, a relative reduction of 11.3% (95% CI [9.2%, 13.3%]), and MAE from .2010 to .1849.

Table 12: Remaining-position conditioning ablation on the common held-out pool. Lower is better.

Metric	Without R_t	With R_t	Paired improvement [95% CI]
Question-balanced Brier	.1130	**.1003**	.0128 [.0099,.0156]
Question-balanced MAE	.2010	**.1849**	.0161 [.0130,.0191]

D.4 Composition and Support Robustness

To test whether the endpoint reversal is created by different system compositions across states, we repeat the comparison inside the TOP5 systems, averaging within system before across systems: the equal-system Search–Commit contrast is $+29.1$ at E_0A_0 (4 systems, 444 trajectories), -2.4 at E_1A_1 (4, 233), -33.8 at E_2A_0 (1, 29), and -7.1 at E_2A_1 (5, 544); only states satisfying the support rule are listed. Restricting instead to systems with supported estimates at both E_0A_0 and E_2A_1, eight systems qualify, seven change from positive to negative, and the mean within-system shift is -30.2 points (system-bootstrap 95% CI [-45.9, -16.2]). Table 13 varies the minimum training-cell support: the endpoint states retain opposite signs throughout the sweep, whereas intermediate states lose support more quickly as the threshold rises and are therefore reported as diagnostic evidence rather than as state laws.

Table 13: Endpoint Search–Commit contrasts under alternative minimum training-cell support.

Minimum support	E_0A_0 (pp)	E_2A_1 (pp)
5	+10.3	-7.5
8	+9.3	-7.6
10	+8.7	-7.4
12	+7.2	-7.3
15	+4.3	-6.9
20	+3.1	-5.4

D.5 Stuck Episodes: Pivot, Repeat, and Escape by System

A trajectory is stuck if it spends five or more consecutive non-terminal turns in E_0A_0 and issues at least five queries; 1,332 of the 5,927 trajectories qualify. Consecutive queries are compared by token Jaccard similarity, with $\geq .85$ counted as a repeat and $\leq .25$ as a pivot; the rates are shares of a trajectory's consecutive query pairs. Escape means terminating in any state with $E \geq E_1$. Table 14 summarizes the three tiers and Table 15 lists every system with at least fifteen stuck trajectories, ordered by accuracy on the stuck subset. Across these thirteen systems, the Spearman correlation between pivot rate and escape rate is .82 and between repeat rate and escape rate $-.86$. The comparison is descriptive and between systems; within a tier, splitting stuck trajectories at the system-wise median pivot rate does not separate escape rates, so the numbers characterize what strong and weak systems do in the same state rather than the effect of pivoting.

Table 14: **How systems behave when stuck without evidence.** Trajectories that spend five or more consecutive turns in E_0A_0 with at least five queries: share of consecutive queries that pivot or repeat the previous query, share that eventually reach gold evidence (escape; 95% question-level bootstrap interval), and final accuracy.

Tier	n	Pivot rate	Repeat rate	Escape (%)	Accuracy
TOP5	353	.706	.005	54.4 [47.5,61.2]	.496
MID	808	.503	.052	23.0 [19.7,26.1]	.189
WEAK	171	.245	.481	0.6 [0.0,1.8]	.006

Table 15: Stuck trajectories by system: count, pivot and repeat rates, escape rate, and final accuracy.

System	Tier	n	Pivot	Repeat	Escape	Accuracy
Claude Opus 5	TOP5	35	.913	.000	.743	.714
Kimi-K2.7-Code	TOP5	103	.646	.010	.621	.573
GLM-5.2	MID	104	.790	.002	.538	.490
GLM-5.1	TOP5	108	.752	.005	.472	.435
Claude Sonnet 4.6	TOP5	59	.751	.001	.475	.424
Kimi-K3	TOP5	48	.527	.003	.479	.396
DeepSeek-V4-Flash	MID	136	.590	.042	.316	.265
WebExplorer-8B-Fix3	MID	175	.428	.017	.234	.183
DeepSeek-V3.2	MID	170	.583	.009	.271	.135
WebSailor-32B	MID	187	.284	.137	.000	.059
R1-Searcher-7B-Official	WEAK	77	.309	.601	.000	.013
DeepSeek-R1	MID	21	.571	.071	.000	.000
WebShaper-32B-Official-MT10K	WEAK	64	.047	.531	.000	.000

E Finding 3: Process-Signal Audit, Scope and Full Results

E.1 Comparison Scope

A prior signal enters the audit only when it can be reconstructed on the logged trajectories without inventing an unavailable quantity and applies to both successful and failed runs (Table 16). Behavior Priming, AgentDiagnose, and AgentProcessBench are re-scored with their official prompts using GPT-4.1 as a common judge; this differs from the judges in the original papers and is held fixed across systems. The audit cohort contains 1,443 correct and 4,484 incorrect trajectories.

Table 16: Eligibility for the common process-signal audit.

Method	Native projection	Audits 1–4	Reason / scope
Log statistics (gold-document recall, counts)	✓	✓	Deterministic retrieval/tool/length statistics available directly from the logs.
Behavior Priming	✓	✓	The four behaviors are reconstructed individually; partial free-text outputs remain missing rather than being set to zero.
AgentDiagnose	✓	✓	The four process competencies only; Objective Quality is excluded because it describes the task rather than the process.
AgentProcessBench	✓	✓	Step labels are averaged to a trajectory-level process score.
TRACE Process Efficiency	×	×	The action-cost mapping required for an exact score is unavailable for our action space.
WebLeaper	×	×	BrowseComp-Plus does not provide the target-entity representation required by the metric.
DAS	×	×	Requires counterfactual search/answer interventions not present in logged trajectories.
AgentRx	×	×	Failure-only labels do not define a common score over correct and incorrect trajectories.
Failure as a Process	×	×	Coding-specific failure annotations do not define a transferable scalar Web-search signal.

E.2 Exact Audit Implementation

Let $Y_{aq} \in \{0, 1\}$ be final correctness and X_{aq} a trajectory-level signal for agent a on question q; all signals, including the binary behavior indicators, are standardized to unit variance, and Y is scaled to percentage points.

Audits 1 and 2. Let $\tilde{X}_{aq} = X_{aq} - \bar{X}_{\cdot q}$ and $\tilde{Y}_{aq} = Y_{aq} - \bar{Y}_{\cdot q}$ be question-demeaned values. Audit 1 reports

$$\beta_1 = \frac{\sum_{a,q} \tilde{X}_{aq} \tilde{Y}_{aq}}{\sum_{a,q} \tilde{X}_{aq}^2}, \tag{8}$$

the OLS slope with question fixed effects, with standard errors clustered by question. Audit 2 replaces the single demeaning by two-way demeaning on additive question and agent effects (iterated until convergence) and reports the same slope β_2 with two-way clustered standard errors by question and agent. The ratio β_2/β_1 is the share of the question-controlled association that survives agent control. Because each agent answers each question once, the two-way design removes additive effects only; an agent-by-question interaction remains in the residual.

Audit 3. For each question q with correct set $\mathcal{P}_q$ and incorrect set $\mathcal{N}_q$,

$$C_Q = \frac{\sum_q \sum_{i \in \mathcal{P}_q} \sum_{j \in \mathcal{N}_q} \left[\mathbb{1}(X_i > X_j) + \frac{1}{2} \mathbb{1}(X_i = X_j) \right]}{\sum_q |\mathcal{P}_q| \, |\mathcal{N}_q|}, \tag{9}$$

the pair-weighted same-question concordance. Confidence intervals and between-signal differences use 10,000 paired bootstrap resamples of questions; all signals are evaluated on the same set of pairs.

Audit 4. After two-way demeaning as in Audit 2, we fit

$$\tilde{Y}_{aq} = \beta_X \tilde{X}_{aq} + \beta_E \tilde{Q}_{\mathrm{EAT},aq} + \varepsilon_{aq} \tag{10}$$

and report β_X ("Prior") and β_E ("EAT"), with the same two-way clustered inference. A prior signal with $\beta_X \approx 0$ next to Q_{EAT} carries no information beyond it; a large β_E next to every prior signal shows that EAT is not explained by any one of them.

Table 17: Complete progressive audit on the 25-system cohort. Audit 1, Audit 2, and Audit 4 are percentage-point changes in correctness per standard deviation of the signal; A2/A1 is the ratio of Audit-2 to Audit-1 association; Audit-4 columns report partial associations of the prior signal and EAT.

Family	Signal	Audit 1	Audit 2	A2/A1	Audit 3 C_Q	Audit 4		n
						Prior	EAT	
Evidence	Gold-document recall	+32.48	+31.91	+0.98	.958	+22.20	+23.10	5927
Interaction	Step Count	+4.48	−8.38	−1.87	–	–	–	5927
Interaction	Search-turn Count	+6.74	−10.93	−1.62	.812	+0.67	+39.16	5927
Behavior Priming	Information Verification	+12.80	+5.61	+0.44	.681	+2.22	+38.27	5737
Behavior Priming	Authority Evaluation	+19.13	+9.88	+0.52	.790	+4.85	+36.92	5724
Behavior Priming	Adaptive Search	+9.51	+1.43	+0.15	.633	+0.13	+39.09	5824
Behavior Priming	Error Recovery	+5.33	−1.73	−0.32	.576	−0.58	+38.85	5791
AgentDiagnose	Backtracking & Exploration	+16.68	+4.21	+0.25	.820	+2.30	+38.55	5899
AgentDiagnose	Task Decomposition	+11.52	+3.20	+0.28	.700	+1.42	+38.67	5927
AgentDiagnose	Observation Reading	+18.46	+10.45	+0.57	.829	+5.15	+36.95	5927
AgentDiagnose	Self-Verification	+17.25	+8.93	+0.52	.809	+4.28	+37.64	5927
AgentProcessBench	Mean Step Label	+13.72	+5.59	+0.41	.747	+2.25	+38.28	5926
EAT	Q_{EAT}	**+32.06**	**+38.83**	**+1.21**	**.961**	–	–	5624

E.3 Equal-Supervision Predictive Comparison

Q_{EAT} is estimated from outcome labels whereas gold-document recall is a direct statistic. To remove this asymmetry, we fit a logistic regression with agent fixed effects under the same five question-level folds (signals standardized on the training folds; no regularization) and evaluate on the common held-out pool of 5,624 trajectories with a supported Q_{EAT}. Recall alone obtains question-balanced Brier error .0753 and same-question concordance .9634; EAT alone obtains .0694 and .9497; the joint model obtains .0585 and .9643. The single-signal ranking therefore depends on the metric, whereas the combination is strongest on both. A broader model with all prior signals gives .0673 (prior only), .0694 (EAT only), and .0555 (joint).

E.4 Agreement Between String-Match Answer Surfacing and the Graders

Checked against the three-grader verdict on all 5,672 autonomous commitments, the string-match A_t has A_1 at commit for 78.8% of graded-correct commitments (91.4% for TOP5, 73.9% for MID, 37.5% for WEAK); 84.0% of A_1 commitments and 7.1% of A_0 commitments are graded correct. A manual audit of every graded-correct E_2A_0 commitment (131 cases) found only literal variants of the gold string: multi-part answers, date and number formats, name forms, and short strings. A normalization that handles these variants raises agreement among graded-correct commitments to 97.2%; all results in this paper use the benchmark's original normalization.