SEQUENTIAL PERMUTATION-BASED EVIDENCE CERTIFICATION FOR ADAPTIVE STOPPING IN LONG-HORIZON SEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon search agents must decide when to stop searching and answer. Stopping too early risks weakly supported answers, while searching after sufficient evidence has been collected wastes computation. Candidate answers evolve during search, and repeated webpages can inflate apparent support without adding new evidence. We introduce **S**equential **P**ermutation-based **E**vidence **C**ertification (SPEC), a training-free controller for frozen search agents. It updates candidate rankings online using the current search context and separately accumulates certification evidence from new sources encountered after each candidate is proposed. It groups repeated pages into source clusters, counting each cluster only once for certification. For candidate comparisons fixed before new evidence arrives, SPEC measures whether each new source provides stronger support than matched retrieval results used as controls. It accumulates this relative support through permutation e-processes, with individual error budgets assigned to dynamically introduced candidates. Certified stopping requires the leading candidate to pass its certification threshold and have a sufficient log-likelihood margin over alternatives. Near the stopping threshold, SPEC checks how much this margin drops when individual sources are removed from a small evidence bundle and raises the required margin when support is concentrated in one source. Under conditional exchangeability of observed and control sources, SPEC bounds the probability of false evidence certification across search steps and dynamically introduced candidates. Experiments on fixed-corpus and open-web search show that SPEC reduces search calls and total cost while maintaining accuracy. Code is available here.

1 INTRODUCTION

Long-horizon search agents interleave querying, browsing, and reasoning to answer questions whose supporting evidence is scattered across documents Yao et al. (2023). Their stopping decisions balance two risks: answering before sufficient evidence is gathered and spending further computation on searches that add little information. Trajectory analyses show that answer accuracy is more closely associated with retrieved evidence than with search volume Liu et al. (2026). Useful evidence often arrives early, yet agents continue through long tails of low-yield retrieval Liu et al. (2026); Soudani et al. (2026). These findings motivate evidence-driven stopping that reduces unnecessary search while maintaining answer quality.

Existing retrieval-control methods use model uncertainty or observed search progress to decide whether further retrieval is warranted. FLARE and DRAGIN use token-level uncertainty to guide retrieval Jiang et al. (2023); Su et al. (2024), while Stop-RAG learns a value-based stopping policy for iterative retrieval Park et al. (2025). In long-horizon web search, RAAC combines document novelty, query diversity, and criteria coverage to choose among continuation, intervention, and stopping Soudani et al. (2026). For answer candidates introduced during search, evidence certification raises a further question: which observations should contribute to a candidate's accumulated support?

Consider an agent that proposes a candidate answer from a webpage and subsequently encounters several pages reproducing content from the same underlying source. Repeated exposure can increase apparent support without adding a new source. Using the initial webpage for certification also

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.

couples the candidate's proposal with its assessment. Even after additional sources are collected, the candidate's advantage over alternatives can still depend heavily on a single source. These cases motivate two complementary checks before stopping: whether a candidate has accumulated support from new sources encountered after its proposal, and how strongly its current advantage depends on individual sources. Assessing this source dependence also incurs computation, motivating selective checks to control their contribution to total system cost.

We introduce **S**equential **P**ermutation-Based **E**vidence **C**ertification (SPEC), a training-free controller for frozen search agents that updates candidate rankings online using the current search context while maintaining separate certification histories. Each history accumulates evidence only from new sources encountered after the candidate is proposed, with repeated pages grouped into source clusters that contribute at most one certification update per candidate. For candidate–opponent comparisons fixed before each agent action, SPEC applies the same scoring function to each eligible new source cluster and matched retrieval results used as controls, then accumulates their relative support through permutation e-processes. Newly introduced candidates receive individual portions of a shared error budget. Under conditional exchangeability of observed and control sources, this construction bounds the probability of false evidence certification across search steps and dynamically introduced candidates. After updating certification and refreshing candidate rankings, SPEC permits certified stopping only when the leading candidate passes its certification threshold and has the required length-normalized log-likelihood margin over alternatives. Near the margin threshold, removing individual sources from a small evidence bundle provides a local source-dependence signal that raises the required margin when support is concentrated in one source. The bounded adjustment allows source-removal checks to be skipped outside this decision band while preserving the outcome of the margin test. Our contributions are as follows:

- We introduce SPEC, a training-free controller for search agents that combines sequential evidence certification with a log-likelihood margin adjusted for local source dependence, keeping online candidate ranking separate from certification.
- We construct source-cluster permutation e-processes with precommitted candidate comparisons and individual error budgets, giving anytime-valid control of false evidence certification for dynamically introduced candidates under conditional exchangeability.
- We evaluate SPEC across multiple frozen agents on a fixed-corpus benchmark and compare its variants in open-web search, assessing answer quality, search calls, total system cost, and certification behavior.

2 Related Work

2.1 Search Agents and Retrieval Control

Search agents combine language-model reasoning with external tools and iterative retrieval. ReAct established the now-standard interleaving of reasoning and action Yao et al. (2023), while Toolformer and WebGPT demonstrated how models can learn or be prompted to call tools and browse the web Schick et al. (2023); Nakano et al. (2021). Retrieval-augmented generation (RAG) connects generation to an external corpus Lewis et al. (2020); later systems improve faithfulness or retrieval adaptivity through self-reflection and corrective retrieval, including Self-RAG and CRAG Asai et al. (2024); Yan et al. (2024). Search-o1 and BrowseComp-Plus extend this setting to reasoning-intensive and deep-research tasks Li et al. (2025); Chen et al. (2026b). A central efficiency question is when further retrieval is useful. Trajectory analyses identify retrieval and utilization gaps and show that useful evidence often arrives early, followed by redundant search Liu et al. (2026); related work describes this late-stage behavior as reasoning stagnation Soudani et al. (2026). FLARE and DRAGIN trigger retrieval from generation uncertainty or information needs Jiang et al. (2023); Su et al. (2024); Adaptive-RAG and Stop-RAG adapt the retrieval depth or learn a value-based stopping policy Jeong et al. (2024); Park et al. (2025). AgentIR and RAAC further use intermediate reasoning states, novelty, and criteria coverage Chen et al. (2026a); Soudani et al. (2026). These methods optimize retrieval behavior, but do not provide an anytime-valid certificate that an adaptively selected candidate is sufficiently supported. PACE provides such testing for accepting self-modifications, rather than for evidence certification of answer candidates Shawn (2026).

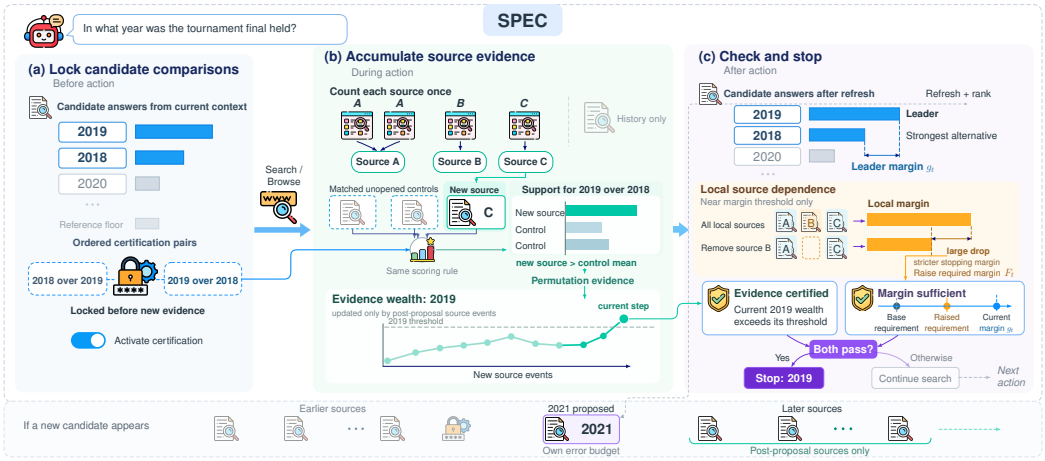


Figure 1: Overview of SPEC. (a) Fix candidate comparisons before each action. (b) Accumulate permutation evidence from post-proposal source clusters against matched controls. (c) Stop when the leader passes certification and a margin test adjusted for local source dependence.

2.2 Sequential Inference for Adaptive Decisions

Sequential inference supports decisions at data-dependent stopping times. E-values and e-processes combine nonnegative evidence while preserving null validity Vovk & Wang (2021); Shafer (2021); Ville's inequality gives an anytime-valid threshold rule, and confidence sequences provide a related uniform-in-time guarantee Howard et al. (2021). Permutation tests and randomization inference offer distribution-light comparisons Welch (1938); Berkson (1942); Chung & Romano (2013), while online multiple-testing methods control error as hypotheses arrive sequentially Benjamini et al. (2006); Ramdas et al. (2017).

3 Method

SPEC maintains a candidate pool with separate ranking and certification channels. The *ranking channel* scores candidates using the search history so far and selects a leader. The *certification channel* accumulates candidate-specific evidence wealth from source clusters encountered after each candidate is proposed, using matched control clusters for comparison. SPEC stops when the leader is certified and passes a log-likelihood-margin gate adjusted for local source dependence. Throughout, q is the question, M the frozen model that both drives the agent and scores answers, and c any rendered context. For a nonempty answer h of token length $|h|$, all scoring uses its length-normalized log-likelihood given q and c:

$$L(h; c) = \frac{1}{|h|} \log p_M(h \mid q, c). \tag{1}$$

3.1 Candidate Pool and Precommitment

After action t the pool $\mathcal{H}_t$ holds at most K concrete short answers; it starts empty. The fixed reference $h_0 = \texttt{insufficient evidence}$ is excluded from $\mathcal{H}_t$ and serves only as a margin floor and abstention label. Let c_t be the complete history after action t. The leader $h_\star$ maximizes $L(h; c_t)$ over $\mathcal{H}_t$, with ranking ties favoring earlier birth. For a fixed comparison pool $\mathcal{H}$, define the margin below; ranking uses $\mathcal{H} = \mathcal{H}_t$:

$$d(h; c) = L(h; c) - \max_{u \in (\mathcal{H} \setminus \{h\}) \cup \{h_0\}} L(u; c), \qquad g_t = d(h_\star; c_t). \tag{2}$$

With a single candidate the maximum runs over h_0 alone; an empty pool has $g_t = -\infty$. Candidates are refreshed after the certification updates of an action whenever (i) a new cluster arrived, (ii) the previous margin was negative, or (iii) m consecutive actions brought neither a new cluster nor a change in the margin. The frozen model then proposes up to K answers from the current history, in a branch that never enters the agent's trajectory; SQuAD-style normalization defines persistent

identities, with empty strings and h_0 discarded. New identities receive consecutive birth indices, assigned lexicographically within a batch. Proposed and active candidates are scored together on c_t and the pool is pruned to K by dropping the lowest scores (ties remove the older). An removed identity keeps its birth index and certification account, so it resumes where it left off if it reappears.

Before each action, SPEC commits to its comparisons using only the available information $\mathcal{F}_{t-1}$: the current top two candidates are selected for certification (each called a *bettor*), each is paired with its strongest concrete opponent, and the previous margin determines the activation gate b_t:

$$\mathcal{A}_t = \mathrm{Top2}(\mathcal{H}_{t-1}), \qquad h'_t(h) = \arg\max_{u \in \mathcal{H}_{t-1} \setminus \{h\}} L(u; c_{t-1}), \qquad b_t = \mathbf{1}\{g_{t-1} \geq \tau_{\mathrm{cert}}\}, \tag{3}$$

with $b_1 = 0$ and $\mathcal{A}_t = \varnothing$ if fewer than two concrete candidates exist. The candidate pairs and the activation gate are fixed before any source cluster of action t is observed. The activation threshold satisfies $\tau_{\mathrm{cert}} \leq \tau$, where τ is the stopping-margin threshold.

3.2 Source-Level Permutation Evidence

SPEC groups pages into *source clusters* using canonical URLs, root domains, original provenance, mirror or repost fingerprints, and near-duplicate body text. Pages assigned to existing clusters update the working history without creating new source events. Each new source cluster defines a *source event* e, processed in arrival order across actions. The focal cluster G_e is compared against K_c *control* clusters drawn by a rule fixed before the episode without consulting certification scores. Controls come first from unopened sibling results (S1), then deeper score&rank-matched results (S2), both from the same query and retriever, and finally from topic-matched results for q (S3). The eligible pool excludes G_e, prior focal clusters (including those from earlier events of the same action), and pages opened by the agent. Focal and control clusters use the same history summary, rendering, passage-selection rule, truncation, prompt position, and scoring function.

Write $x_0 = G_e$ for the focal cluster, $x_1, \ldots, x_{K_c}$ for the controls, and σ_e for the summary of the history before e (it includes the earlier events of the same action). For each bettor h and its precommitted opponent $h' = h'_t(h)$, write f_e for the pair-specific scoring function; (σ_e, x) denotes the concatenation of the summary and rendered cluster:

$$f_e(x) = \frac{e^{L(h;\sigma_e,x)}}{e^{L(h;\sigma_e,x)} + e^{L(h';\sigma_e,x)}}, \qquad E_e^h = \frac{(K_c+1)\, f_e(x_0)}{\sum_{k=0}^{K_c} f_e(x_k)}. \tag{4}$$

The contrastive score measures support for h relative to h'. The factor E_e^h exceeds one exactly when the focal score exceeds the mean control score.

An event updates h only if $b_t = 1$, $h \in \mathcal{A}_t$, and the event follows its proposal. If fewer than K_c distinct controls are available, the event is skipped for all bettors. A check invariant to permutations of the score tuple skips only the affected candidate's comparison if the tuple is nonfinite or the normalizer underflows. Skipped comparisons use $E_e^h = 1$ and are logged. All scores are evaluated in the log domain, and each cluster contributes at most one update per candidate.

3.3 Sequential Certification

Each candidate carries one wealth stream per betting fraction $\lambda \in \Lambda$, a fixed nonempty finite subset of $(0, 1]$. All streams are initialized to one at birth, and certification begins at the next event, keeping the evidence used to propose a candidate outside its certification history. The streams are updated multiplicatively at each event and averaged with equal weights:

$$W_{e,\lambda}^h = W_{e^-,\lambda}^h \big(1 - \lambda + \lambda E_e^h\big), \qquad W_t^h = \tfrac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} W_{t,\lambda}^h, \tag{5}$$

where e^- denotes the instant before event e, after all earlier events have been processed. The state $W_{t,\lambda}^h$ is recorded at the end of action t; initialization applies to new identities, and existing streams remain unchanged when no event updates them.

For a shared error budget $\alpha \in (0, 1)$, the j-th proposed identity receives capital $w_h = 1/(j(j+1))$ and error budget αw_h; the capitals of all identities ever born sum to at most one, removed ones included. A candidate is *certified* at action t when $w_h W_t^h \geq 1/\alpha$. Reentry restores the birth index, capital, and every wealth stream without applying missed updates. Opponent changes preserve wealth, and certification is re-evaluated after every action.

3.4 Local Source Dependence and the Dual-Gate Stop

The margin gate accounts for concentrated dependence on one source cluster. When cluster G first arrives, SPEC records its gain $\Delta g(G) = \max_{h\in\mathcal{H}_{t-1}} d(h; c_G^+) - \max_{h\in\mathcal{H}_{t-1}} d(h; c_G^-)$. This is the leader margin change over the pre-refresh pool $\mathcal{H}_{t-1}$ (used to evaluate d) between the complete histories immediately before (c_G^-) and after (c_G^+) arrival; an empty maximum is $-\infty$. Recorded gains remain fixed; a non-finite gain counts as zero only for bundle selection. Up to J clusters with the largest gains form the bundle set B_t, with ties favoring earlier arrival.

For a set B of source clusters, $c(B)$ renders their stored text bundles in arrival order under the common question q; removal preserves the remaining bundles' content and order. With the full-history leader $h_\star$, candidate pool, and scoring procedure fixed, SPEC rescores $\mathcal{H}_t \cup \{h_0\}$ in small local contexts with all bundles and with each bundle removed, measuring the largest relative drop:

$$a_t = d\big(h_\star; c(B_t)\big), \qquad a_t^{-G} = d\big(h_\star; c(B_t\backslash\{G\})\big), \qquad F_t = \min\Big\{1,\ \frac{\max_{G\in B_t}[\,a_t - a_t^{-G}\,]_+}{|a_t| + \varepsilon}\Big\}, \tag{6}$$

where $[\cdot]_+$ denotes the positive part, $\varepsilon > 0$, and an empty B_t gives $F_t = 0$. The baseline and ablated margins are evaluated in local contexts with the leader, candidate pool, and scoring procedure fixed.

For $\kappa > 0$, the margin threshold rises from τ to $\tau + \log(1+\kappa)$ as F_t increases from zero to one. Both gates must pass:

$$\mathrm{STOP}_t = \mathbf{1}\{|\mathcal{H}_t| \geq 2\} \cdot \mathbf{1}\{w_{h_\star} W_t^{h_\star} \geq 1/\alpha\} \cdot \mathbf{1}\{g_t \geq \tau + \log(1 + \kappa F_t)\}. \tag{7}$$

An empty or singleton pool returns $\mathrm{STOP}_t = 0$ without evaluating the remaining factors. For larger pools, $F_t \in [0,1]$ makes the final factor zero when $g_t < \tau$ and one when $g_t \geq \tau + \log(1+\kappa)$. These branches use g_t alone; only the intermediate band requires F_t, giving the same margin decision as evaluating it at every step. A non-finite required score fails the corresponding gate. When $\mathrm{STOP}_t = 1$, SPEC returns the leader as a *certified stop*, including at the action cap.

If m consecutive actions bring neither a new cluster nor a margin change and SPEC has not stopped, it selects the source cluster with the largest nonnegative local margin drop; earlier arrival breaks ties, including when all drops are zero. Target selection can run outside the stopping band. With a concrete leader, a nonempty bundle set, and remaining budgets, the model extracts the selected source's key assertion and drafts a query for independent corroboration. Ordinary agent behavior continues if extraction or local scoring fails. After the next precommitment, the query replaces one host action, consuming one step of $T_{\max}$ and one unit of the corroboration budget B_{corr}; results follow the usual source-clustering and certification steps. Drafting leaves wealth unchanged; at most one query is pending, and exhausting either budget prevents scheduling.

At $T_{\max}$, certified stopping takes priority. If the leader has a concrete opponent and passes certification but fails the margin gate, SPEC returns a *certified low-margin answer*. Otherwise, the fixed output policy returns a nonempty pool's leader as an *uncertified answer* when an answer is required, or *abstains* when allowed; an empty pool always abstains.

3.5 Conditional Certification Guarantee

Certification is meant to be *safe* in one precise sense: a candidate whose retrieved sources carry no more support than matched controls should rarely become certified, at any time and however many candidates appear. We state the condition under which this holds.

Assumption 1 (Conditional exchangeability)**.** For a candidate h and a event e, let $\mathcal{F}_{e^-}$ be the information available before the focal/control assignment and the scores at e are revealed (earlier events and wealth, the precommitted pair and gate, σ_e, the retrieval pools, and the fixed eligibility rules). A comparison is *admitted* if $b_t = 1$, $h \in \mathcal{A}_t$, e follows the birth of h, and K_c distinct controls exist. The null H_0^h holds if, at every admitted event, $(f_e(x_0), \ldots, f_e(x_{K_c}))$ is exchangeable given $\mathcal{F}_{e^-}$.

Theorem 1 (Candidate-level certification control)**.** *Let $\mathcal{N}_0$ be the set of candidate identities satisfying H_0^h. For every $\alpha \in (0,1)$,*

$$\Pr\big(\exists\, h \in \mathcal{N}_0,\ \exists\, t:\ w_h W_t^h \geq 1/\alpha\big) \leq \alpha. \tag{8}$$

Table 1: Main results on BrowseComp-Plus. We report accuracy and gold recall (%), and search calls as the median (mean) per question. Macro denotes the average over the five agents, and bold indicates the best method per column, with fewer search calls counted as better.

Method	GPT-oss 20B-high	Qwen3.5 35B-A3B	Tongyi-DR 30B-A3B	DeepSeek V4 Pro	GLM 5.2	Macro
Accuracy / gold recall (%)						
Native	45.2 / 55.7	54.9 / 62.6	49.6 / 60.2	70.1 / 76.6	78.2 / 79.9	59.6 / 67.0
Fixed-Cap-20	46.6 / 55.5	54.6 / 62.5	49.5 / 60.2	71.2 / 76.5	77.5 / 79.9	59.9 / 66.9
Fixed-Cap-40	44.8 / 55.7	54.9 / 62.6	50.4 / 60.2	69.8 / 76.6	77.7 / 79.9	59.5 / 67.0
Full SPEC	**50.4 / 58.8**	**59.0 / 64.2**	**51.2 / 61.9**	**71.3 / 77.1**	**79.3 / 80.6**	**62.2 / 68.5**
Search calls, median (mean)						
Native	25 (29.5)	20 (22.9)	26.5 (35.9)	26 (28.7)	16 (19.1)	22.7 (27.2)
Fixed-Cap-20	20 (**16.7**)	20 (15.7)	20 (**16.3**)	20 (17.2)	16 (14.2)	19.2 (**16.0**)
Fixed-Cap-40	25 (24.8)	20 (21.3)	26.5 (25.4)	26 (25.6)	16 (17.7)	22.7 (23.0)
Full SPEC	**17** (18.8)	**14 (14.8)**	**15** (21.3)	**14 (15.7)**	**12 (14.1)**	**14.4** (16.9)

Theorem 1 is what makes *certified* a calibrated word: over an entire run, every candidate the pool ever holds, and every action at which SPEC might stop, the probability of certifying a candidate on evidence no more informative than its controls is at most α, with no fixed horizon or hypothesis set. It also explains the design: precommitting the pair and the gate fixes the conditioning information $\mathcal{F}_{e^-}$, symmetric scoring of focal and controls makes exchangeability attainable, and post-birth wealth with capital w_h is exactly what the supermartingale and union-bound argument in the appendix needs. Assumption 1 is the single interface to the retrieval process: given everything known before a page is opened, the chosen page must not be systematically privileged over its unopened siblings in separating h from h'. Matched controls are designed to make this plausible and the negative-control experiments to test it; certification is thus a statement about accumulated evidence relative to that protocol (see the Appendix A.9).

4 Experiments

4.1 Experimental Setup

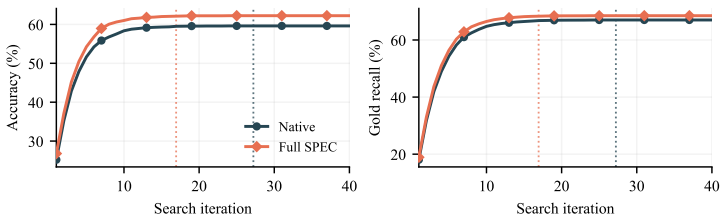
Benchmarks. All headline experiments use BrowseComp-Plus Chen et al. (2026b): 830 English questions over a fixed ∼100K-document corpus with human-annotated evidence and gold qrels. We adopt the evaluation and implementation protocol of Liu et al. (2026) in full: a shared ReAct harness Yao et al. (2023) exposing `search` and `visit` tools over the official prebuilt Qwen3-Embedding-8B index (top-5, 512-token snippets), a 128-turn / 150-minute outer cap, and a frozen `gpt-4o-2024-08-06` judge. Holding the corpus, retriever, and judge fixed isolates the effect of the stopping controller; intermediate-behavior metrics follow the definitions of Soudani et al. (2026). Open-web generalization uses three further benchmarks (§4.6).

Agents and methods. We evaluate five frozen agents—GPT-oss-20B-high, Tongyi-DR-30B-A3B, Qwen3.5-35B-A3B, DeepSeek V4 Pro, and GLM 5.2—and, for a controlled comparison, run every method within the same agent, checkpoint, prompts, and question order. Baselines comprise (i) the *Native* agent, which stops on its own, and (ii) *Fixed-Cap-20/40*, which forces an answer at the 20th/40th search episode; ablations add the reduced variants *Margin*, *Margin+Fragility*, and *Margin+E-process*. Full SPEC uses $K{=}4$, $K_c{=}2$ with the S1→S2→S3 control ladder, $\Lambda{=}\{0.25, 0.5, 1\}$, $\alpha{=}0.1$, $\kappa{=}3$, $J{=}3$, $m{=}5$, a corroboration budget $B_{\mathrm{corr}}{=}2$, and 512-token bundles, with the preregistered stopping threshold $\tau{=}1.0$ and its pre-activation margin $\tau_{\mathrm{cert}}{=}\tau{-}0.5{=}0.5$ shared by all five agents. Query-altering configurations are physically rerun, the remainder are replayed offline from logged states, and the headline configuration is verified online against its replay.

Evaluation metrics. We report accuracy, gold recall, and search calls as defined by Liu et al. (2026), together with total cost, which sums agent, search, controller, and judge cost under a frozen price table (USD); inference uses $B{=}10{,}000$ question-level paired bootstrap.

Table 2: Total cost per question in USD, which sums agent, search, controller, and judge cost under a frozen price table. Ratio is the cost of Native relative to Full SPEC, and per-method statistics are provided in Appendix A.6.

Agent	Native	Cap-20	Cap-40	Full SPEC	Ratio
GPT-oss-20B-high	0.452	0.263	0.383	0.350	1.29
Qwen3.5-35B-A3B	0.351	0.244	0.327	0.276	1.27
Tongyi-DR-30B-A3B	0.506	0.239	0.363	0.370	1.37
DeepSeek V4 Pro	0.425	0.263	0.381	0.291	1.46
GLM 5.2	0.243	0.185	0.227	0.229	1.06
Macro Average	0.395	0.239	0.336	0.303	1.30



(a) Intermediate accuracy (b) Cumulative gold recall

Figure 2: Intermediate accuracy (a) and cumulative gold recall (b) of the native agent and Full SPEC over search iterations, macro-averaged over agents, with dotted lines marking each method's mean stopping iteration. The protocol follows Soudani et al. (2026).

4.2 Main Results on BrowseComp-Plus

SPEC maintains answer quality while searching less. Across all five agents, Full SPEC improves accuracy over the native agent and both fixed caps, with the largest gains on the weakest agents (Table 1); gold recall moves in the same direction, so the search reduction does not come at the expense of surfaced gold evidence. Mean search calls fall by roughly a quarter to nearly a half, and the bootstrap intervals for the search-call differences exclude zero for every agent. The accuracy differences are likewise positive for every agent, though their intervals exclude zero only on the weakest agent, where the gain is largest, so we read the accuracy evidence as a consistent direction rather than as per-agent significance.

Accuracy gains are not explained by early forcing. Fixed-Cap-20 also removes the wasted tail but leaves accuracy at roughly the native level, as truncation alone neither helps nor hurts on average. The improvements specific to SPEC therefore track its certification and fragility gates rather than merely stopping earlier, a decomposition §4.5 quantifies.

Total cost drops once the controller is charged for. With the controller's candidate scoring, control-source scoring, and bundle ablations all included, Full SPEC is cheaper than native for every agent (Table 2); the Native/Full ratio exceeds the preregistered 1.2 efficiency threshold for four of five agents, the exception being the already-frugal GLM 5.2, where SPEC remains cost-neutral while improving accuracy. Fixed-Cap-20 is cheaper still but concedes the accuracy gains, so SPEC occupies the accuracy–cost frontier among the compared methods.

4.3 Trajectory-Level Search Analysis

SPEC stops near evidence saturation, not at a fixed budget. Figure 2 plots intermediate accuracy and cumulative gold recall for the native agent and Full SPEC; the fixed caps replay the logged native trajectories, so their curves coincide with the native ones before each cap and are omitted. SPEC's recall curve sits slightly above the native one, because its corroboration queries surface gold a little faster. SPEC's mean stopping point sits where the recall curve flattens, whereas native agents continue through a long low-yield region, and by its stopping point SPEC has already recovered slightly more gold evidence than the native run recovers by its own later stop, consistent with Table 1.

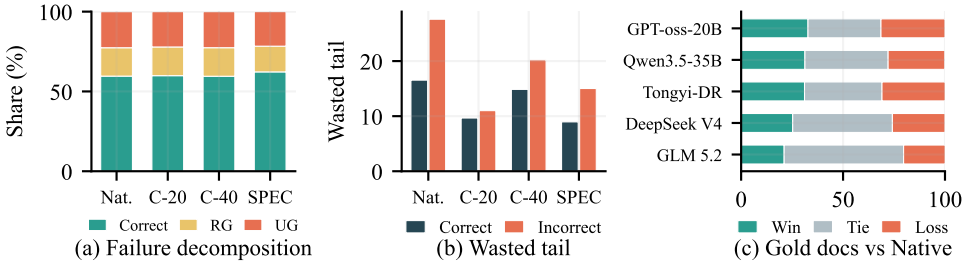


Figure 3: Trajectory-level effects of stopping control, macro-averaged over agents. Panel (a) shows the shares of questions that are correct or exhibit a retrieval gap (RG) or a utilization gap (UG), and panel (b) shows the mean wasted tail in episodes on correct and incorrect trajectories. Panel (c) reports the win/tie/loss shares of Full SPEC against native in retrieved gold documents. Nat./C-20/C-40/SPEC in (a) and (b) abbreviate Native, Fixed-Cap-20/40, and Full SPEC.

Table 3: Validity of sequential certification at α=0.1. Crossing is the null-sequence crossing rate with a Wald 95% CI, and Sel. gap is the difference between agent-selected and randomized focal sources in percentage points with a cluster bootstrap 95% CI. No family is rejected by the exact one-sided binomial test at level 0.05.

Null family	n	Crossing	95% CI	Sel. gap	95% CI
Exact null (synthetic, pooled)	10,000	0.092	[0.086, 0.098]	—	—
Wrong candidate vs. gold	300	0.057	[0.031, 0.083]	+0.9	[−1.2, +3.1]
Matched-difficulty wrong pair	300	0.047	[0.023, 0.071]	−0.4	[−2.6, +1.8]
Candidate label permutation	300	0.083	[0.052, 0.115]	+0.4	[−1.6, +2.3]
Irrelevant placebo candidate	300	0.040	[0.018, 0.062]	−0.3	[−2.5, +1.9]

The removed search is mostly the wasted tail. Following Liu et al. (2026), an episode is *productive* if it surfaces new evidence documents, and the *wasted tail* is the episodes after the last productive one. Figure 3b shows that incorrect native trajectories run 1.3 to 2.3 times the wasted tail of correct ones across agents, 1.7 times macro-averaged, reproducing the diagnosis that motivated evidence-driven stopping; SPEC compresses the tail on both outcome classes. At the episode level (Appendix A.4), SPEC raises the productive share of episodes while leaving visit precision and first-gold-hit positions essentially unchanged, indicating that the controller prunes low-yield continuation rather than cutting evidence gathering short.

Error structure shifts toward the retrieval side, and per-question evidence is preserved. In Figure 3a, macro-averaged over agents, SPEC's accuracy gain appears as a small decrease in both retrieval- and utilization-gap errors; per agent, the retrieval gap falls on four of five agents and the utilization gap on four of five, with the two exceptions roughly flat, so certified stopping does not convert retrieval failures into utilization failures. Figure 3c compares final retrieved gold documents per question against the same agent's native run: wins exceed losses for every agent and ties dominate, confirming that the search reduction is not paid for by lost gold evidence on individual questions.

4.4 Validity of Sequential Certification

Certification is only meaningful if Theorem 1's guarantee survives implementation and deployment; we test this in three layers (Table 3). On exactly exchangeable synthetic sequences, the pooled crossing rate over ten thousand runs stays below α, so the implementation does not inflate the test. On four *operational* nulls built from real trajectories—a wrong candidate against the gold answer, a matched-difficulty wrong pair, permuted candidate labels, and an irrelevant placebo—no family's point estimate exceeds α and no family is rejected by the exact one-sided binomial test at level 0.05; the upper limit of the label-permutation interval (0.115) sits slightly above α, so for this family we cannot rule out a crossing rate modestly above the nominal level, and we flag it as the one to monitor rather than claim it as a pass. Finally, comparing certification outcomes when the focal source is the one the agent actually selected against a randomized choice from the same eligible pool yields small selection-gap estimates whose intervals cover zero in all families, so agent page selection does not detectably privilege the focal side.

Table 4: Core component ablation on Qwen3.5-35B-A3B. Cert. is certification coverage in percent, $\text{Wrong}_{\text{cert}}$ is the fraction of wrong answers among certified stops in percent, and bold marks the best value per column, with lower being better for Search and Cost.

Variant	Acc	Gold Rec.	Search (med.)	Cost	Cert	$\text{Wrong}_{\text{cert}}$
Margin	52.5	60.4	**10**	**0.219**	—	—
Margin+Fragility	54.8	62.4	11	0.237	—	—
Margin+E-process	55.2	61.7	13	0.256	**99.6**	44.6
Full SPEC	**59.0**	**64.2**	14	0.276	99.2	**40.5**

Table 5: Open-web generalization with a single frozen agent (Qwen3.5-35B-A3B), evaluated on BrowseComp Open Web with 200 questions and one run, and on EvoBrowseComp EN/ZH with 400 questions each and means over 3 runs. ER is the fraction of runs hitting the official tool-call cap in percent, and bold indicates the best accuracy per column.

	BrowseComp Open Web			EvoBrowseComp-EN			EvoBrowseComp-ZH		
Variant	Acc	Search	Cost	Acc	Search	ER	Acc	Search	ER
Margin	46.0	14.4	0.287	48.2	11.9	19.7	40.8	11.6	19.4
Margin+Fragility	48.5	15.5	0.311	50.6	12.9	12.1	43.1	12.5	11.4
Margin+E-process	49.0	17.1	0.335	51.0	14.3	11.5	43.4	13.8	10.5
Full SPEC	**53.0**	18.5	0.362	**54.8**	15.4	11.9	**47.3**	15.0	12.7

Certification behavior at task time. Across the five agents, the controller certifies a stopping candidate on the vast majority of questions, with median certification episodes between 12 and 17; the fraction of wrong answers among certified stops tracks each agent's overall error rate rather than exceeding it, and control-source exhaustion stays below one event in ten (Appendix A.5). We emphasize the reading: certification is a statement about accumulated evidence relative to matched controls, not about factual correctness, and the wrong-among-certified rates should be read jointly with the crossing tests above.

4.5 Ablation Studies

Table 4 isolates the two certification channels on Qwen3.5-35B-A3B. The margin gate alone stops earliest and cheapest but *loses* accuracy relative to native, matching the diagnosis that a stable-looking leader margin is not sufficient evidence. Each channel adds accuracy on top of the margin gate, and their combination recovers the full headline gain; the margin-only variant's failure to preserve accuracy is what rules out the simple-stopping explanation. Because certification fires nearly regardless of eventual correctness (Appendix A.5), the e-process channel is best understood not as a correctness discriminator but as a minimum-evidence requirement, preventing the margin gate from committing until corroborating support has accumulated across independent source clusters, which is exactly the failure mode of the margin-only variant. Finer-grained ablations (Appendix A.3) show that removing source clustering roughly doubles the null crossing rate, resetting wealth on comparator switches is the most destructive single change, and the bundle-level fragility signal agrees with a full-context reference on most band states at a fraction of its cost.

4.6 Open-Web Generalization

The fixed corpus provides qrels-grounded diagnostics but no page dynamics, so on the open web we fix a single agent and compare only the SPEC family (Table 5). The component ordering from the ablation reproduces in all three settings: each certification channel adds accuracy over the margin gate, and Full SPEC is best throughout. Two open-web specifics stand out. On BrowseComp Open Web the variants spend *more* search calls than on the fixed corpus: without qrels-side redundancy and with noisier rankings, the certification gates request additional corroboration before stopping. On EvoBrowseComp, the full controller sharply reduces the fraction of runs exhausting the official tool-call budget relative to the margin-only variant, so the gains on evolving knowledge come from better-directed rather than longer search (per-run results in Appendix A.7).

AI USE STATEMENT

In this work, generative AI tools based on large language models were used only for language editing and polishing of the manuscript. All technical content, experiments, and reported numbers were produced solely by the authors through the evaluation pipeline described in the paper. All AI-polished text was checked against the underlying experimental records. No AI tool was used for developing code, running experiments, generating data, or any task involving human subjects or personal data, and no AI tool replaced peer review or author approval of the final submission. We take responsibility for the final content of this work, including any text polished with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or personally identifiable information. All benchmarks, corpora, and models are either publicly available or accessed through standard commercial APIs under their terms of service, and all evaluation data are derived from these public resources. Model outputs are used solely to answer factual benchmark questions, and no content generation targets individuals or protected groups. We report the full computational cost of every method under a fixed price table. The efficiency gains studied here reduce rather than increase the resource footprint of deployed search agents. We also note that our certification mechanism is a statement about accumulated evidence relative to matched controls, not about factual correctness. The paper repeatedly draws this distinction to avoid over-claiming the reliability of certified answers.

REPRODUCIBILITY STATEMENT

All experiments follow a frozen, preregistered protocol in which the benchmark, corpus, retriever, judge, agent checkpoints, controller hyperparameters, stopping thresholds τ, and price table are fixed in advance and fully specified in Section 4.1. Query-altering configurations are physically rerun, the remainder are replayed from logged states, and the headline configuration is verified online against its replay. All statistical comparisons use question-level paired bootstrap with $B{=}10{,}000$ resamples and fixed seeds. The appendix provides the complete algorithm, per-agent results, per-run open-web outcomes, component ablations, the controller cost breakdown, and the exact binomial acceptance region used in the validity tests. The evaluation code, configuration files, per-question outputs, and the data package used to generate every table and figure will be made publicly available upon publication.

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

Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

Joseph Berkson. Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37(219):325–335, 1942. doi: 10.1080/01621459.1942.10501760.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Jimmy Lin, Akari Asai, and Victor Zhong. Agentir: Reasoning-aware retrieval for deep research agents. *arXiv preprint arXiv:2603.04384*, 2026a. doi: 10.48550/arXiv.2603.04384. URL https://arxiv.org/abs/2603.04384.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. Browsecomp-plus: A fair and disentangled evaluation benchmark for deep search agents. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 22349–22370.

Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.ACL-LONG.1023. URL https://doi.org/10.18653/v1/2026.acl-long.1023.

EunYi Chung and Joseph P. Romano. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507, 2013. doi: 10.1214/13-AOS1090.

Etienne Gauthier, Francis R. Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction. *CoRR*, abs/2503.13050, 2025. doi: 10.48550/ARXIV.2503.13050. URL https://doi.org/10.48550/arXiv.2503.13050.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. 2020.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Zhengbao Jiang, Frank F. 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/.

Adrian Kieback, Uyiosa Philip Amadasun, Aman Chadha, and Aaron Elkins. TASR: training-free adaptive stopping for iterative retrieval. *CoRR*, abs/2606.13814, 2026. doi: 10.48550/ARXIV.2606.13814. URL https://doi.org/10.48550/arXiv.2606.13814.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

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

Qi Liu, Jiaxin Mao, Fengbin Zhu, and Tat-Seng Chua. Diagnosing search behavior and failure modes in long-horizon search agents. *CoRR*, abs/2608.01913, 2026. doi: 10.48550/ARXIV.2608.01913. URL https://doi.org/10.48550/arXiv.2608.01913.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332, 2021. URL https://arxiv.org/abs/2112.09332.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-rag: Value-based retrieval control for iterative RAG. *CoRR*, abs/2510.14337, 2025. doi: 10.48550/ARXIV.2510.14337. URL https://doi.org/10.48550/arXiv.2510.14337.

Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, and Michael I. Jordan. Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems 30*, pp. 5650–5659, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7f018eb7b301a66658931cb8a93fd6e8-Abstract.html.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *CoRR*, abs/2210.01948, 2022. doi: 10.48550/ARXIV.2210.01948. URL https://doi.org/10.48550/arXiv.2210.01948.

Daeyoung Roh and Donghee Han. HALT: verification-aware stopping for retrieval-augmented search agents. *CoRR*, abs/2608.02009, 2026. doi: 10.48550/ARXIV.2608.02009. URL https://doi.org/10.48550/arXiv.2608.02009.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021. doi: 10.1111/rssa.12647. URL https://doi.org/10.1111/rssa.12647.

Zayx Shawn. Pace: Anytime-valid acceptance tests for self-evolving agents. *arXiv preprint arXiv:2606.08106*, 2026. doi: 10.48550/arXiv.2606.08106. URL https://arxiv.org/abs/2606.08106.

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. doi: 10.48550/arXiv.2608.15191. 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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12991–13013. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.702. URL https://doi.org/10.18653/v1/2024.acl-long.702.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020. URL https://doi.org/10.1214/20-AOS2020.

B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362, 1938. doi: 10.1093/biomet/29.3-4.350.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024. doi: 10.48550/arXiv.2401.15884.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

APPENDIX

A.1 THE PROCESS OF SPEC ALGORITHM

Algorithm 1 One controller iteration of SPEC for action t

Precommit from $\mathcal{F}_{t-1}$: bettors $\mathcal{A}_t$, opponents $h'_t(h)$, gate b_t (Eq. 3).
The frozen agent executes one action, replaced by a pending corroboration query if present; each new source cluster defines an event e.
for each source event e of action t **do**
  Record the cluster's gain Δg; finalize the eligible control pool.
  if $b_t = 1$, $\mathcal{A}_t \neq \varnothing$, and K_c distinct controls exist **then**
    Draw K_c distinct controls. For each $h \in \mathcal{A}_t$: score focal and controls symmetrically, form E_e^h (Eq. 4), and update wealth (Eq. 5).
  else
    Leave all wealth unchanged.
  end if
end for
Refresh candidates if triggered; new identities receive capital w_h, initialize every wealth stream to one, and become eligible at the next event.
Rank $\mathcal{H}_t \cup \{h_0\}$ on c_t; compute g_t (Eq. 2) and evaluate the stopping gates as described after Eq. 7.
Stop and return a *certified stop* if $\mathrm{STOP}_t = 1$ (Eq. 7); otherwise apply the cap rule at $T_{\max}$, or continue and schedule corroboration when eligible.

A.2 ADAPTED STOPPING REFERENCES

Table 6 compares Full SPEC against two adapted stopping references on Qwen3.5-35B-A3B. TASR-style (Kieback et al., 2026) stops when two consecutive intermediate answers coincide under out-of-fold isotonic calibration; HALT-style (Roh & Han, 2026) stops when generated claims are all matched by the released verifier. Both are fixed adaptations, not native BrowseComp implementations, and neither trains on the test questions. Each stops earlier than native at native-level accuracy, but neither recovers the accuracy gain of the certified controller, ruling out answer stability or claim coverage alone as an explanation of SPEC's improvements.

Table 6: Adapted stopping references on Qwen3.5-35B-A3B, where Search reports search calls as median (mean) and Inc is the incomplete rate in percent.

Method	Acc	Gold Rec.	Search	Inc	Cost	Adaptation
Native	54.9	62.6	20 (22.9)	0.6	0.351	native harness
Fixed-Cap-20	54.6	62.5	20 (15.7)	0.0	0.244	forced answer at episode 20
Fixed-Cap-40	54.9	62.6	20 (21.3)	0.0	0.327	forced answer at episode 40
TASR-style	54.8	62.3	16 (13.5)	0.6	0.211	out-of-fold isotonic calibration
HALT-style	55.4	62.4	18 (14.6)	0.6	0.229	generated claims + released verifier
Full SPEC	**59.0**	**64.2**	14 (14.8)	0.6	0.276	full frozen method

A.3 DETAILED ABLATIONS

All ablations in this subsection run on Qwen3.5-35B-A3B over the full 830 questions. Null crossing rates reuse the operational-null construction of §4.4 on the same questions.

Control sources and source clustering (Table 7). Removing source clustering roughly doubles the operational null crossing rate and costs about two points of accuracy: repeated pages inflate apparent support, exactly the failure mode the clustering layer is designed to remove. Among control layers, the full S1→S2→S3 ladder is best on task metrics; S1-only degrades certification coverage because eligible sibling controls are frequently exhausted, and the C3 hard-sensitivity alternative trades coverage for a slightly lower crossing rate. Doubling the number of controls (K_c=4) improves coverage and crossing but increases total cost beyond the full method's, so the headline uses K_c=2.

Table 7: Control-source and clustering ablations on Qwen3.5-35B-A3B, where Search reports mean search calls per question, Cert is certification coverage in percent, NullX is the operational null crossing rate, and Kc is the number of controls per event. C1 and C3 denote sensitivity-only control layers excluded from the headline ladder, namely uniformly sampled corpus sources and hard near-duplicate sources from different origins, respectively.

Variant	Controls	Kc	Acc	Rec	Search	Cert	NullX	Cost
Full SPEC	S1→S2→S3	2	**59.0**	**64.2**	14.8	99.2	0.072	0.276
No clustering	S1→S2→S3	2	57.0	62.6	16.9	94.2	0.158	0.296
S1 only	S1	2	56.6	62.1	17.3	92.7	0.122	0.257
S2 control	S2	2	58.1	63.4	15.8	96.6	0.088	0.264
S3 control	S3	2	57.5	63.4	16.4	94.9	0.089	0.261
C1 uniform	C1	2	58.2	63.7	15.8	96.8	0.075	0.268
C3 hard	C3	2	57.8	63.2	14.1	100.0	0.059	0.282
Four controls	S1→S2→S3	4	59.3	64.4	15.5	100.0	0.054	0.356

Wealth, candidates, and comparators (Table 8). The three-stream mixture $\Lambda = \{0.25, 0.5, 1\}$ outperforms every single betting fraction, consistent with the mixture hedging against unknown evidence-growth regimes. Betting only on the top-1 candidate reduces certification coverage: without a second bettor, a late-rising candidate has no accumulated wealth. Resetting wealth on comparator switches is the most destructive variant, confirming that the precommitted, switch-tolerant wealth accounting is doing real work; a fixed numeric h_0 floor performs close to the literal floor, so the margin gate is not an artifact of the specific abstention string.

Table 8: Wealth, candidate, and comparator ablations on Qwen3.5-35B-A3B, where Search reports mean search calls per question and $\text{Wrong}_{\text{cert}}$ is the fraction of wrong answers among certified stops in percent.

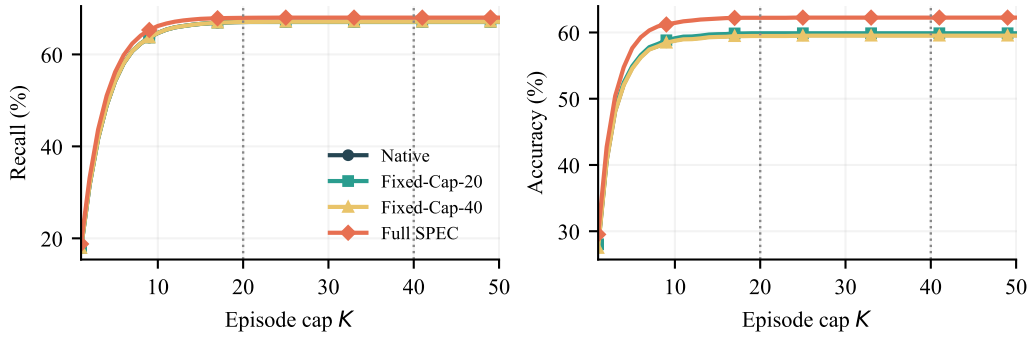
Variant	Streams	Targets	Acc	Search	Cert	$\text{Wrong}_{\text{cert}}$	Cost
Full SPEC	mixture	top-2	**59.0**	14.8	99.2	40.5	0.276
λ=0.25 only	0.25	top-2	57.5	15.9	94.7	41.2	0.287
λ=0.5 only	0.5	top-2	58.3	15.3	97.4	40.6	0.280
λ=1 only	1	top-2	57.8	13.9	96.8	40.1	0.268
Top-1 betting	mixture	top-1	57.8	13.7	92.0	41.0	0.273
Reset on switch	mixture	top-2	56.8	16.7	92.0	41.5	0.287
Fixed numeric floor	mixture	top-2	58.3	15.0	97.4	41.4	0.274

Fragility and targeted verification (Table 9). The bundle-level fragility signal agrees with a full-context reference computation on about nine in ten band states while costing a fraction of its compute; forcing the full-context variant everywhere yields no additional accuracy. Shrinking the bundle (J=2) or weakening the correction (κ=1) degrades the false-stop rate smoothly, and κ=10 overcorrects at higher cost, supporting the preregistered κ=3, J=3. Removing targeted corroboration costs about one point of accuracy, and its false-stop rate rises, matching its role in resolving concentrated support before the gates commit.

Table 9: Fragility and targeted-verification ablations on Qwen3.5-35B-A3B, where Search reports mean search calls per question, Agr is the agreement with the full-context reference, and FS is the false-stop rate in percent.

Variant	Context	J	κ	Tgt	Acc	Search	Agr	FS	Cost
Full SPEC	bundle	3	3	on	59.0	14.8	0.94	22.4	0.276
No fragility	none	0	0	on	55.2	13.7	—	31.1	0.256
Full-context fragility	full	3	3	on	59.2	14.8	1.00	21.2	0.344
Bundle J=2	bundle	2	3	on	58.7	15.1	0.93	25.2	0.269
Bundle κ=1	bundle	3	1	on	58.2	15.5	0.90	26.2	0.268
Bundle κ=10	bundle	3	10	on	58.8	14.9	0.96	20.3	0.292
No targeted corroboration	bundle	3	3	off	58.0	14.3	0.94	28.5	0.258

A.4 TRAJECTORY-LEVEL ANALYSES



(a) Padded cumulative recall

(b) Upper-bound accuracy

Figure 4: Budget-side upper bounds, macro-averaged over agents. Panel (a) shows padded cumulative recall and panel (b) shows upper-bound accuracy $\widehat{\mathrm{Acc}}_K$ under an episode cap K, with dotted lines marking $K{=}20$ and $K{=}40$. The protocol follows Liu et al. (2026).

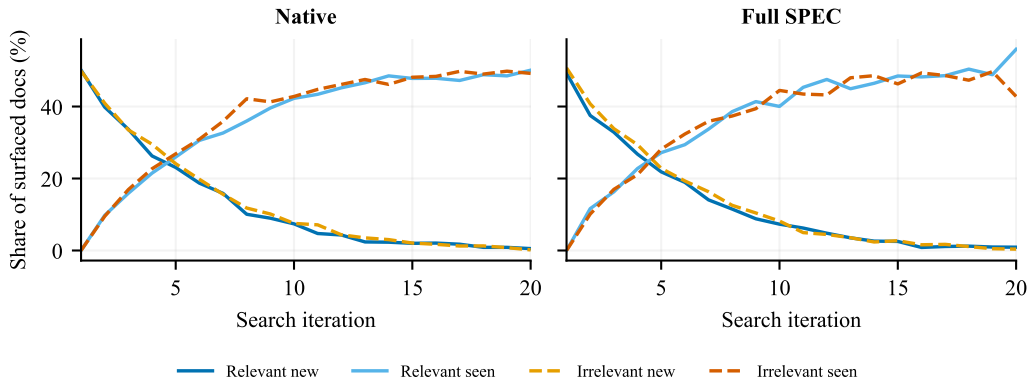


Figure 5: Retrieved-document dynamics over the first 20 iterations for native (left) and Full SPEC (right), macro-averaged over agents, showing the share of surfaced documents that are relevant or irrelevant and new or seen. Definitions follow Soudani et al. (2026).

Table 10 summarizes the trajectory diagnostics per agent and method. Under SPEC, the productive share of episodes rises and the wasted tail roughly halves on both outcome classes, while first-gold-hit positions, visit precision, and the retrieval/utilization split move little—the controller removes continuation after saturation rather than changing where evidence is found. Repeat-query and malformed-call rates both drop because trajectories end before degenerate late-stage query behavior emerges. Figure 4 shows the budget-side recall and accuracy upper bounds: a $K{=}20$ cap retains most of each agent's accuracy, which is why Fixed-Cap-20 is a strong cost baseline and why SPEC's accuracy gains must come from the gates rather than the cap. Figure 5 confirms that the search SPEC removes is dominated by seen and irrelevant documents rather than relevant-new ones, and Figure 6 shows the corresponding reduction in final context, concentrated in search snippets.

A.5 TASK-TIME CERTIFICATION BEHAVIOR

Table 11 reports the per-scenario exact-null breakdown behind the pooled test in the main text. Table 12 shows that certification coverage is high for every agent and that the wrong-among-certified rate tracks each agent's base error rate; the certification gates fire nearly regardless of eventual correctness, which is expected: the wealth process tests relative evidence, not truth, and the margin

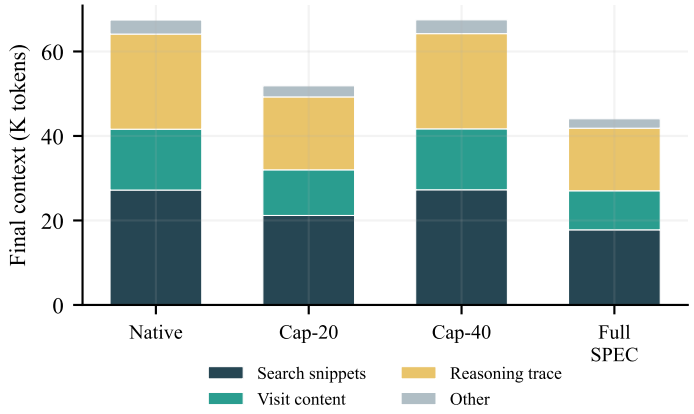


Figure 6: Final context composition by token source, macro-averaged over agents.

Table 10: Episode, retrieval, and query summary on BrowseComp-Plus, where Prod, Red, and Unprod are the shares of productive, redundant, and unproductive episodes in percent, ZeroP is the share of questions with no productive episode, and Tail is the wasted-tail episode count as median (mean). FirstGold is the first gold-hit episode as median (mean), NoGold is the share of questions never surfacing gold, RG and UG are the retrieval and utilization gaps in percent, VP is visit precision in percent, and Repeat and Bad are the shares of repeated and malformed queries in percent.

Agent	Method	Prod	Red	Unprod	ZeroP	Tail	FirstGold	NoGold	RG	UG	VP	Repeat	Bad
GPT-oss-20B	Native	7.5	27.1	65.4	10.4	18 (24.0)	2 (3.1)	28.0	27.8	27.0	54.2	6.4	8.7
GPT-oss-20B	Full SPEC	12.5	25.3	62.2	7.5	10 (13.1)	2 (2.9)	22.9	22.4	27.2	54.4	4.0	5.2
Qwen3.5-35B	Native	11.0	26.4	62.6	6.7	13 (17.1)	2 (2.9)	21.6	21.1	24.0	56.2	6.4	9.9
Qwen3.5-35B	Full SPEC	16.0	24.2	59.9	8.3	8 (9.8)	2 (2.6)	20.7	19.3	21.7	56.3	3.1	5.6
Tongyi-DR-30B	Native	7.3	30.3	62.3	4.0	20 (30.1)	2 (2.5)	21.9	21.3	29.0	43.1	6.0	8.5
Tongyi-DR-30B	Full SPEC	12.2	27.1	60.7	4.7	10 (15.9)	2 (2.7)	20.5	20.2	28.6	43.1	3.6	5.7
DeepSeek V4 Pro	Native	10.4	29.7	59.9	2.9	19 (22.6)	2 (2.6)	11.3	11.0	18.9	63.1	5.5	8.4
DeepSeek V4 Pro	Full SPEC	18.9	26.7	54.4	1.7	8 (9.9)	2 (2.4)	11.4	10.6	18.1	61.6	3.7	5.8
GLM 5.2	Native	16.1	28.1	55.8	1.0	10 (13.5)	2 (2.2)	9.2	7.7	14.1	68.3	5.9	8.4
GLM 5.2	Full SPEC	21.1	25.9	52.9	1.8	6 (8.8)	2 (2.3)	9.6	8.1	12.7	67.6	4.0	5.3

gate carries the accuracy signal. Table 13 shows candidate recall saturates at the pool size $K{=}4$ (the R@8 diagnostic adds little), gold candidates are typically born early enough to accumulate wealth, and comparator switches average fewer than two per trajectory, so the composite-null concern of frequent opponent changes rarely binds in practice.

A.6 System and Cost Profile

Table 14 expands the cost analysis of the main text. SPEC reduces visits and final context for every agent, with the largest context savings on the snippet-heavy Tongyi-DR, whose overflow-driven incompletes also drop. Table 15 breaks the controller's own cost into components: candidate refreshes and focal/control scoring dominate, while bundle fragility is cheap precisely because it runs only inside the decision band, and targeted corroboration is rare by construction.

A.7 Open-Web Details

On BrowseComp Open Web (Table 16), the accuracy ordering of the main text is accompanied by a monotone reduction in judge-based retrieval gaps, while utilization gaps stay flat—consistent with the fixed-corpus finding that certification-driven stopping mostly rescues questions whose evidence had not yet been gathered. Per-run EvoBrowseComp results (Table 17) show the variant ordering is stable across runs, with run-to-run variation well below the between-variant gaps.

Table 11: Exact-null implementation test at α=0.1, reporting the crossing rate over 2,500 strictly exchangeable synthetic sequences per scenario. The pooled count of 919 lies below the rejection threshold $X \geq 1{,}051$ of the exact one-sided binomial test of $H_0 : p \leq \alpha$ at level 0.05.

Scenario	Sequences	Crossing count	Crossing rate
Static candidates, three wealth streams	2,500	217	0.087
Dynamic birth, α-spending	2,500	221	0.088
Predictable gate, skipped steps (E=1)	2,500	240	0.096
Full frozen state machine	2,500	241	0.096
Pooled	10,000	919	0.092

Table 12: Task-time certification behavior per agent under Full SPEC, where Coverage is the certified-stop share in percent, P(cert|·) is the certification rate on correct and wrong answers in percent, $\text{Wrong}_{\text{cert}}$ is the share of wrong answers among certified stops in percent, and Cert ep. is the median certification episode. Exhaust is the share of events without eligible controls in percent.

Agent	Coverage	P(cert|corr.)	P(cert|wrong)	$\text{Wrong}_{\text{cert}}$	Cert ep.	Exhaust
GPT-oss-20B	98.0	100.0	95.9	48.6	17	9.5
Qwen3.5-35B	99.2	100.0	97.9	40.5	13	8.8
Tongyi-DR-30B	95.4	100.0	90.6	46.3	14.5	9.7
DeepSeek V4 Pro	99.6	99.8	99.2	28.5	14	9.4
GLM 5.2	98.7	99.8	94.2	19.8	12	8.8

A.8 Signal Informativeness

Table 18 scores the controller's internal signals as predictors of whether an iteration surfaces new evidence. The continuous signals are signed by construction: the margin compares the leader against the abstention floor and its strongest opponent, so it drifts below zero when evidence stops arriving, and although the wealth multiplier has conditional mean one under the null, its realizations fall below one on uninformative iterations, making mean log wealth negative by Jensen's inequality. The bundle response R_t likewise keeps the sign of the source-removal effect, since removing a bundle member can help as well as hurt; the stop rule clips it into the nonnegative fragility F_t of Eq. 6. Among these signals, log certification wealth is by far the most informative, which is why the e-process rather than the raw margin serves as the accumulation channel; the margin is informative but weaker, and the bundle response is a risk signal rather than a productivity one, consistent with its role in the dual-gate stop.

Table 13: Candidate and termination diagnostics under Full SPEC, where R@k is candidate recall in percent at pool size k, GoldBirth is the gold-candidate birth step as median (mean), Switch is the number of comparator switches per trajectory, and Band and Bet are band entries and admitted bets per trajectory. R@8 is reported as a diagnostic only because the pool size remains $K{=}4$, and the termination columns sum to 100% up to rounding. The certified low-margin and abstention categories of the cap-time policy occur on 0.0% of questions for every agent and are omitted.

Agent	R@1	R@2	R@4	R@8	GoldBirth	Switch	Band	Bet	Exhaust	Cert. stop	Uncert.	MaxTurns
GPT-oss-20B	53.0	64.9	72.8	73.5	8.5 (9.5)	1.81	5.0	10.0	9.5	98.0	0.1	1.9
Qwen3.5-35B	58.4	69.8	77.0	77.2	7 (7.9)	1.51	4.1	8.1	8.8	99.2	0.2	0.6
Tongyi-DR-30B	55.2	66.0	71.9	72.5	7.0 (9.8)	2.19	5.2	10.9	9.7	95.4	0.0	4.6
DeepSeek V4 Pro	68.2	81.1	85.9	85.9	7 (8.2)	1.56	4.3	8.6	9.4	99.6	0.2	0.1
GLM 5.2	72.4	83.1	86.5	86.6	6 (7.2)	1.47	3.8	7.7	8.8	98.7	0.8	0.5

Table 14: System and cost profile on BrowseComp-Plus, where Inc is the incomplete rate in percent, Visit is the number of visit calls as median (mean), Ctx is the final context in thousands of tokens as median (mean), Q/turn is the number of queries per turn, and Ovhd and Cost are the controller and total cost per question in USD.

Agent	Method	Inc	Visit	Ctx	Q/turn	Ovhd	Cost
GPT-oss-20B	Native	1.8	10 (10.2)	80.2 (93.2)	1.01	0	0.452
	Cap-20	0.0	6 (5.9)	58.8 (54.5)	1.01	0	0.263
	Cap-40	0.0	8 (8.7)	80.2 (79.0)	1.01	0	0.383
	Full SPEC	1.9	5 (6.6)	53.7 (60.5)	1.01	0.057	0.350
Qwen3.5-35B	Native	0.6	6 (8.3)	62.4 (72.2)	1.73	0	0.351
	Cap-20	0.0	6 (5.7)	55.3 (50.2)	1.74	0	0.244
	Cap-40	0.0	6 (7.7)	62.4 (67.2)	1.73	0	0.327
	Full SPEC	0.6	5 (5.4)	42.4 (47.4)	1.74	0.046	0.276
Tongyi-DR-30B	Native	6.5	31 (41.8)	75.4 (101.0)	1.01	0	0.506
	Cap-20	0.0	21 (19.0)	51.9 (48.2)	1.01	0	0.239
	Cap-40	0.0	31 (29.6)	75.4 (72.8)	1.01	0	0.363
	Full SPEC	4.6	18 (24.9)	46.0 (61.7)	1.01	0.063	0.370
DeepSeek V4 Pro	Native	0.0	11 (11.5)	80.3 (86.6)	1.49	0	0.425
	Cap-20	0.0	7 (6.9)	57.0 (54.0)	1.49	0	0.263
	Cap-40	0.0	10 (10.3)	80.3 (77.7)	1.49	0	0.381
	Full SPEC	0.1	5 (6.4)	46.9 (49.9)	1.49	0.048	0.291
GLM 5.2	Native	0.1	6 (6.7)	39.1 (46.0)	1.87	0	0.243
	Cap-20	0.0	5 (5.0)	36.3 (35.3)	1.87	0	0.185
	Cap-40	0.0	6 (6.2)	39.1 (43.1)	1.87	0	0.227
	Full SPEC	0.5	5 (5.1)	31.2 (35.5)	1.86	0.044	0.229
Macro	Native	1.8	12.8 (15.7)	67.5 (79.8)	1.42	0	0.395
	Cap-20	0.0	9.0 (8.5)	51.9 (48.4)	1.42	0	0.239
	Cap-40	0.0	12.2 (12.5)	67.5 (68.0)	1.42	0	0.336
	Full SPEC	1.5	7.6 (9.7)	44.0 (51.0)	1.42	0.052	0.303

Table 15: Controller cost profiler for Full SPEC on Qwen3.5-35B-A3B, where Calls is the number of invocations per trajectory, GPU-s and Wall are seconds per trajectory, USD is the cost per question, and Share is the share of controller cost in percent.

Component	Calls	GPU-s	Wall (s)	USD	Share
Candidate refresh generation	14.9	0.80	0.92	0.0138	30%
Candidate teacher-forcing	59.4	3.21	3.69	0.0064	14%
Focal/control source scoring	8.1	0.44	0.51	0.0101	22%
Wealth/state update	14.9	0.80	0.92	0.0014	3%
Bundle fragility	4.1	0.22	0.25	0.0041	9%
Full-context fragility reference	0.6	0.07	0.08	0.0028	6%
Targeted verification query	0.8	0.04	0.05	0.0073	16%
Controller total	102.8	5.58	6.42	0.0459	100%

Table 16: BrowseComp Open-Web results on Qwen3.5-35B-A3B over 200 questions, where Coverage is the LLM-judged evidence sufficiency in percent, RG and UG are the judge-based gaps, Search and Visit are mean calls, and Cost is the cost per question in USD.

Variant	Acc	Coverage	Search	Visit	RG	UG	Cost
Margin	46.0	58.5	14.4	6.0	30.5	23.5	0.287
Margin+Fragility	48.5	60.0	15.5	6.5	28.0	23.5	0.311
Margin+E-process	49.0	59.5	17.1	7.2	28.0	23.0	0.335
Full SPEC	**53.0**	**62.0**	18.5	7.8	24.0	23.0	0.362

Table 17: EvoBrowseComp per-run results on Qwen3.5-35B-A3B with 400 questions per run and 3 runs per variant, where all variants of a language are evaluated over the same three-day window. Acc 1–3 are per-run accuracies in percent, and Mean is the run mean.

Language	Variant	Acc 1	Acc 2	Acc 3	Mean
English	Margin	47.75	48.75	48.25	48.2
English	Margin+Fragility	50.0	51.0	50.75	50.6
English	Margin+E-process	50.5	51.5	51.0	51.0
English	Full SPEC	54.25	55.25	54.75	54.8
Chinese	Margin	40.75	40.25	41.5	40.8
Chinese	Margin+Fragility	43.0	42.5	43.75	43.1
Chinese	Margin+E-process	43.25	42.75	44.25	43.4
Chinese	Full SPEC	47.25	46.75	48.0	47.3

Table 18: Controller signals as predictors of whether a search iteration is productive (Full SPEC runs). Prod/Non-prod: signal means on productive and non-productive iterations. Continuous signals are signed by construction; $R_t = (a_t - a_t^{-G})/(|a_t| + \varepsilon)$ is the bundle response on band states, which the stop rule clips to the fragility F_t of Eq. 6.

Signal	Direction	AUROC	Prod	Non-prod
g_t (margin)	higher	0.779	0.62	-0.15
Δg_t (current-source gain)	higher	0.677	0.38	-0.08
$\log(w_h W_t^h)$ (evidence wealth)	higher	**0.958**	1.42	-0.31
R_t (signed bundle response)	raw sign	0.599	0.19	-0.06
MarginPass (binary)	binary	0.663	0.71	0.29
Cert (binary)	binary	0.626	0.66	0.34

A.9 PROOFS

Notation is that of Section 3; source events are totally ordered across actions. Recall that $E_e^h = 1$ whenever a comparison is not admitted or the permutation-invariant guard fires, that admission is decided from $\mathcal{F}_{e^-}$, and that the guard is a symmetric function of the unordered score tuple.

A.9.1 PER-EVENT SYMMETRY

Lemma 1. Fix h and an admitted event e at which H_0^h holds. Then $\mathbb{E}\big[E_e^h \mid \mathcal{F}_{e^-}\big] = 1$, and $E_e^h \in (0, K_c + 1)$ when the guard does not fire.

Proof. Abbreviate $f_k = f_e(x_k)$ and $S = \sum_{k=0}^{K_c} f_k$, so that by Eq. 4

$$E_e^h = (K_c + 1)\,\frac{f_0}{S}, \qquad f_k \in (0,1) \;\Rightarrow\; 0 < \frac{f_0}{S} < 1 \;\Rightarrow\; E_e^h \in (0, K_c + 1). \tag{9}$$

Under H_0^h, $(f_0, \ldots, f_{K_c})$ is exchangeable given $\mathcal{F}_{e^-}$ and S is symmetric, hence

$$\mathbb{E}\big[f_k/S \mid \mathcal{F}_{e^-}\big] = \mathbb{E}\big[f_0/S \mid \mathcal{F}_{e^-}\big], \qquad k = 0, \ldots, K_c. \tag{10}$$

Summing over k,

$$(K_c + 1)\,\mathbb{E}\big[f_0/S \mid \mathcal{F}_{e^-}\big] = \sum_{k=0}^{K_c} \mathbb{E}\big[f_k/S \mid \mathcal{F}_{e^-}\big] = \mathbb{E}\big[S/S \mid \mathcal{F}_{e^-}\big] = 1. \tag{11}$$

Given $\mathcal{F}_{e^-}$ and the unordered tuple, the focal label is uniform on $\{0, \ldots, K_c\}$ and independent of the (symmetric) guard event; the display above holds off that event and $E_e^h = 1$ on it, so $\mathbb{E}[E_e^h \mid \mathcal{F}_{e^-}] = 1$.

The statistic $f_0 \big/ \big(\frac{1}{K_c+1} S\big)$ is the conformal e-value of Gauthier et al. (2025), with the focal cluster as test point and the controls as calibration set.

A.9.2 CLOSURE OF THE WEALTH

Lemma 2. Fix h with H_0^h and let τ_h be its birth event, i.e., the first event after the refresh that proposed h. For every $\lambda \in \Lambda$ the stream $(W_{e,\lambda}^h)_{e \geq \tau_h}$ of Eq. 5, and its equal-weight mixture $W_e^h = \frac{1}{|\Lambda|} \sum_\lambda W_{e,\lambda}^h$, are nonnegative supermartingales with respect to $(\mathcal{F}_{e^-})_e$ with $W_{\tau_h}^h = 1$.

Proof. Admission is $\mathcal{F}_{e^-}$-measurable; on a non-admitted event the factor equals 1, and on an admitted event Lemma 1 gives

$$\mathbb{E}\big[1 - \lambda + \lambda E_e^h \mid \mathcal{F}_{e^-}\big] = 1 - \lambda + \lambda \cdot 1 = 1, \qquad 1 - \lambda + \lambda E_e^h \;\geq\; 1 - \lambda \;\geq\; 0. \tag{12}$$

Since $W_{e^-,\lambda}^h \geq 0$ is $\mathcal{F}_{e^-}$-measurable, Eq. 5 yields

$$\mathbb{E}\big[W_{e,\lambda}^h \mid \mathcal{F}_{e^-}\big] = W_{e^-,\lambda}^h\, \mathbb{E}\big[1 - \lambda + \lambda E_e^h \mid \mathcal{F}_{e^-}\big] = W_{e^-,\lambda}^h. \tag{13}$$

A fixed convex combination of nonnegative supermartingales is a nonnegative supermartingale (Vovk & Wang, 2021; Ramdas et al., 2022), and all streams start at 1 at τ_h:

$$W_{\tau_h}^h = \tfrac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} 1 = 1. \tag{14}$$

A.9.3 PROOF OF THEOREM 1

Proof. Fix $h \in \mathcal{N}_0$, let $J(h)$ be its birth index so that $w_h = 1/\big(J(h)(J(h)+1)\big)$, and set $c_h = 1/(\alpha w_h)$. The birth event τ_h is a stopping time and $J(h)$ is assigned at τ_h, so c_h is $\mathcal{F}_{\tau_h}$-measurable. By Lemma 2 under the conditional law given $\mathcal{F}_{\tau_h}$ (Howard et al., 2020, Lemma 1),

$$\Pr\Big(\sup_{e \geq \tau_h} W_e^h \geq c_h \;\Big|\; \mathcal{F}_{\tau_h}\Big) \;\leq\; \frac{W_{\tau_h}^h}{c_h} \;=\; \alpha\, w_h \qquad \text{on } \{\tau_h < \infty\}. \tag{15}$$

Since W_t^h is W_e^h at the last event of action t, $\{\exists t:\ w_h W_t^h \geq 1/\alpha\} \subseteq \{\sup_{e \geq \tau_h} W_e^h \geq c_h\}$; taking expectations,

$$\Pr\big(\exists t:\ w_h W_t^h \geq 1/\alpha\big) \;\leq\; \alpha\, \mathbb{E}\big[w_h\, \mathbf{1}\{\tau_h < \infty\}\big]. \tag{16}$$

Summing over $h \in \mathcal{N}_0$ (union bound, then Tonelli),

$$\Pr\big(\exists h \in \mathcal{N}_0,\ \exists t:\ w_h W_t^h \geq 1/\alpha\big) \;\leq\; \alpha\, \mathbb{E}\Big[\sum_{h \in \mathcal{N}_0:\ \tau_h < \infty} w_h\Big]. \tag{17}$$

Distinct identities have distinct birth indices, so almost surely

$$\sum_{h \in \mathcal{N}_0:\ \tau_h < \infty} w_h \;\leq\; \sum_{j \geq 1} \frac{1}{j(j+1)} \;=\; \sum_{j \geq 1} \Big(\frac{1}{j} - \frac{1}{j+1}\Big) \;=\; 1. \tag{18}$$

Combining the last two displays gives Eq. 8. Random birth order is absorbed by conditioning on $\mathcal{F}_{\tau_h}$ before taking expectations; no independence across identities is used.