The model stops at the first answer it can confidently defend to you. Not when it's right.

Background to the post of the same name, drawn from the working paper How Do We Use This?

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

Preprint. Not peer reviewed. Reproduced from the working paper as revised 21 September 2026 and corrected 23 and 25 September 2026.

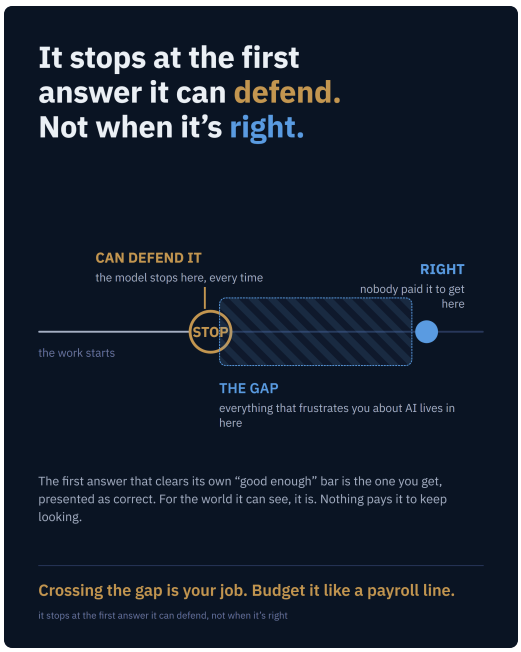
IN PLAIN TERMS

An AI agent ends its work at the first answer it can defend against what it has already gathered, not at the point where the answer is correct. For the limited evidence in front of it, that answer usually is defensible, which is why it arrives sounding confident. Nothing in how these systems are trained rewards the extra search that would expose a mistake.

In the practice studied, about 48% of the operator's messages carried correction or redirection, roughly a third after adjusting for the detector's error. Published work finds the same shape: agents that commit early and defend it, answers left unverified even when they happen to be right, and positions that move under pressure whether or not the pressure is correct.

What helps is a check from outside the agent, anchored on evidence, delivered at the moment it stops.

About this paper. This is one of seven short background papers written to go with a series of plain-English posts. Every section below is reproduced word for word from the working paper *How Do We Use This?* (Atkinson, 2026; revised 21 September 2026, corrected 23 and 25 September 2026), and keeps that paper's section, figure and table numbers, so a reference to a section or figure not reproduced here resolves in the full paper. Only the plain-terms summary, the post's figure, and the short labels that say which section each passage comes from were written for this part. The full paper is linked from every post in the series.



The post's figure. The agent stops where its answer can be defended; the correct answer lies further on, and the gap between them is where the misses live. A teaching summary of §1 and Figure 1; it plots no quantity.

FROM §1 · THE MECHANISM

A working agent receives a task, gathers evidence, and at some point stops gathering and answers. This paper is about what decides that point. The field note that started this program put it plainly:

> *"[The model is] trained to present as fact the first answer, the only criteria for that answer being that it thinks it meets the user's expectations and is defensible based on the available information... Correctness of the answer is not a factor."*
>
> The author's research notes, 30 August 2026

Restated as a mechanism: the stopping rule is satisfiability, not correctness. The search ends when the agent can build a defensible account from what it already holds. Because further evidence can only make an account harder to defend, the marginal incentive is to stop. Written as two stopping times, the defect is the distance between them.

$$t^{\star} = \min \{ t : D(a_t \mid E_t) \} \quad \text{the moment the agent stops}$$

$$t^{\circ} = \min \{ t : C(a_t) \} \quad \text{the moment the task is satisfied}$$

E_t is the evidence held at step t, which grows only by searching. D is defensibility against E_t; C is correctness. D can hold while C fails, and the set of turns where it does is exactly where the misses live. Nothing in the objective rewards closing the distance.

The incentive language above is an as-if description, in the sense Wang and Huang (2026) give their own model: it says the behavior can be rationalized this way, not that any incentive is represented inside the model, whose next step is simply the likeliest continuation of a context that already reads as resolved. The defect is therefore located at the termination decision rather than in the reasoning that precedes it, which is why interventions aimed at reasoning quality do not reach it. Figure 1 draws the mechanism as a position on one axis.

Independent work points the same way. Mehta (2026) describes premature commitment in tool-using agents: runs settle on one reading of the evidence early and then defend it, and a hidden-state signature of that settling can be read at runtime. The same work reports that the signature does not separate agents that settled on the right answer from agents that settled on the wrong one, a result its author grades as suggestive rather than confirmed, and it advises deployers to send settled cases to an external verifier or a human rather than resample. A detector of the stop is not a corrector of it.

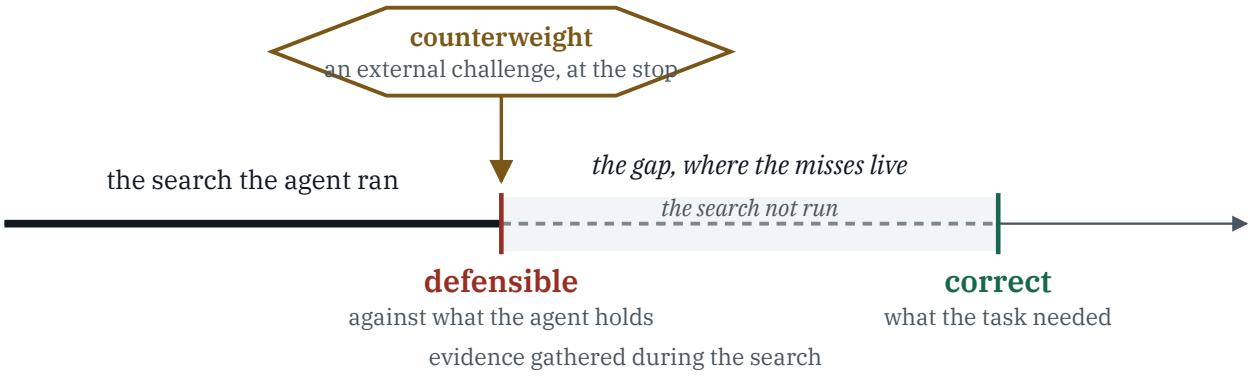


Figure 1. The stop problem as a position on one axis. The agent stops where its answer becomes defensible against what it holds; the correct answer lies further along. The counterweight (§8) is an external challenge timed to the stop. Schematic only: no quantity is plotted, and the counterweight's challenge engine is implemented but not deployed. **Schematic.**

FROM §2.1 · AN OLD PROBLEM WITH A NEW SUBJECT

2.1 The problem's lineage

The stop problem is not new; only its subject is. A search that ends when the searcher is satisfied rather than when the answer is right is Simon's satisficing (Simon, 1955), and when people stop gathering information has its own literature (Browne, Pitts & Wetherbe, 2007). In clinical reasoning the same failure is named premature closure, the tendency to settle on a diagnosis before it has been verified, and it is a documented source of diagnostic error (Graber, Franklin & Gordon, 2005). The standard remedy there is also named: a cognitive forcing function, a deliberate prompt that interrupts the fluent stop and forces a second look (Croskerry, 2003), studied more recently as a way to reduce a person's

overreliance on an AI's suggestions (Buçinca, Malaya & Gajos, 2021). The counterweight of §8 is a cognitive forcing function delivered to a machine.

Three results from the language-model literature bound what such a check can be, and each was read in full. Models do not reliably repair their own reasoning without an outside signal: asked to self-correct with no external feedback, they can make correct answers worse and need an oracle to gain anything (Huang et al., 2023). Challenge is double-edged: across ten models on seven tasks, a bare "are you sure?" flipped answers 46% of the time and cost about 17% of accuracy (Laban et al., 2023), which is why the counterweight must discriminate on error rather than merely apply pressure. And redundancy among similar checkers does not buy independence: independently written versions of a program fail together far more often than independence would predict (Knight & Leveson, 1986), the reason a deciding body must differ in kind from the bodies it decides between (§6.5). The design rule of §8.3 restates, for a practice run by AI agents, Goodhart's and Campbell's laws (Campbell, 1979) and the software-inspection requirement of a non-author reviewer (Fagan, 1976).

The stop problem in LLM agents is also, as of 2026, an active literature of its own, and this paper's own search did not find it until late. Roh and Han (2026) build HALT, an external verification layer that halts a frozen search agent once retrieved evidence covers every reasoning hop. What it saves depends sharply on what it is given, and the distinction is worth carrying: supplied with gold claims, a diagnostic setting, it cuts search loops by 20 to 45% across three datasets; supplied with claims the system generates for itself, which is the deployable setting, the cuts fall to between 2 and 18%. Answer accuracy passes formal non-inferiority testing in both. Run in the other direction, forcing continuation on the trajectories HALT flags as under-covered, it raises population exact match by 8.2 points on HotpotQA and by 2.7 and 4.3 points on the other two datasets. Ko et al. (2026) study what they call illusory completion, where an agent believes a task resolved while constraints remain unverified, and measure underverified-answer rates ranging from 52.1% for the strongest system tested to 76.3% for a ReAct baseline, with human annotators agreeing at Fleiss $\kappa = 0.74$. The finding this program should sit up for is the one inside the successes: even among answers that were *factually correct*, 19.1% were still underverified on that strongest system. That is the defensible-versus-correct distinction measured directly, and measured on answers that a correctness-only scorer would have recorded as wins. Their inference-time tracker, LiveLedger, reduces underverified answers by up to 26.5% and raises accuracy by up to 11.6%. Xu et al. (2026) gate the commitment itself: their execution layer refuses an agent's edit or patch until the evidence the task requires has actually been observed, worth 4.8 to 11.8 points of Pass@1 across 500 SWE-bench Verified instances while cutting token use by up to 12.1%. In the same comparison, bolting a post-hoc self-review step onto the baseline agent instead *lowered* Pass@1, by 1.4 and 1.8 points on the two models tested, which the authors read as confirming that post hoc self-review cannot recover from decisions made on insufficient evidence. Luo et al. (2026) treat stopping as abstention, find that agents abstain late when they abstain at all, and improve timely abstention from 26.7% to 57.4% by distilling past trajectories into stopping rules injected into the agent's context.

Behind those four sits an older and larger line that this program's searches repeatedly missed, because it is indexed under retrieval rather than under stopping. Adaptive-retrieval controllers have been deciding when an agent should stop gathering for several years, using reflection-token critics, model uncertainty, or question complexity to route the decision; Roh and Han (2026, §2) survey the family.

The clearest statement of its premise is Park, Cho and Lee (2025), who cast iterative retrieval as a finite-horizon Markov decision process, learn a value-based controller for when to stop, and conclude that adaptive stopping is a missing component of current agentic systems. A second field asks the same question in the vocabulary of abstention rather than search: selective prediction and learning to defer, where Piatrashyn et al. (2026) hand a decision to a larger model once a small model's calibrated uncertainty crosses a threshold, and report that deferring a modest fraction of decisions matches the expensive model's quality. These literatures do not talk to each other, and we measured it rather than asserting it. We read the complete reference list of fourteen papers: six on retrieval stopping, five on agent termination, and three on deferral. Citations to the foundational reject-option and selective-prediction lineage appear in three of the fourteen, and all three are in the deferral group. None of the five agent-termination papers cites any of it. Five of the six retrieval-stopping papers cite none of it either; the sixth cites selective classification and then explicitly declines to inherit its guarantee, writing that it uses the same conceptual separation but claims no formal risk guarantee, and it is the only paper of the fourteen that calls for the bridge to be built. Between the retrieval-stopping group and the deferral group there are eighteen possible citation pairs in each direction, and the realized count is zero both ways; between agent termination and deferral it is zero of fifteen each way. Each cluster cites itself densely, so this is not a sparse field. It is three internally connected and mutually hermetic islands, each working on when a system should stop. A fourth and older line asks the same question of a network's internal computation rather than of its search, and it sits outside the fourteen-paper count: adaptive computation, in which a recurrent network learns how many computational steps to take between receiving an input and emitting an output (Graves, 2016).

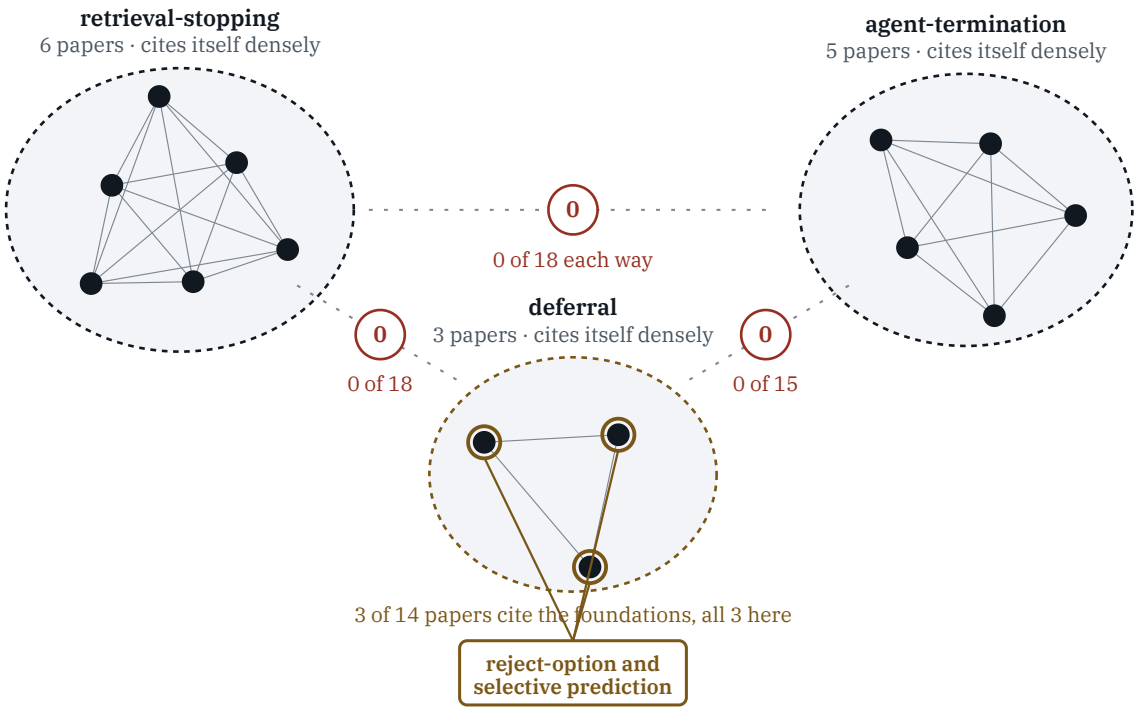


Figure 2. The 2026 stopping literature as a citation graph, counted rather than asserted (§2.1). Each cluster cites itself densely; the crossings that would connect them are empty, zero of eighteen possible pairs each way between retrieval-stopping and deferral and zero of fifteen between agent-termination and deferral, and only three of the fourteen papers reach the fifty-year-old selective-prediction and learning-to-defer foundations. Three internally connected, mutually hermetic islands. **Measured.**

We report that as a finding rather than a complaint, and it implicates this paper directly. A search entered from any one of those vocabularies returns a small set that looks far more novel than it is, which is exactly what happened here across several passes before the citation graphs were walked. The mechanism is the paper's own thesis wearing a bibliography: a search stops when it has enough to write a defensible related-work section.

Three consequences follow, and none of them is comfortable. First, the counterweight's clauses (a) to (c) describe a design that others have now built and measured, mostly with positive results, so this program should be read as a fourth instance rather than a proposal. Second, Xu et al.'s comparison is direct evidence for the design rule of §8.3: gating before commitment beat correcting after it. Third, Luo et al.'s reusable stopping rules, distilled from past trajectories and injected without updating any model parameters, are H1's mechanism working in a setting where the rules came from trajectories rather than from an operator. Against that background this paper makes no claim to priority over any of it, and §7.2 records why that kind of claim was removed rather than narrowed.

3 P1 The missing weight is correctness

The objectives these systems are trained and deployed under reward an answer the user accepts, and nothing in them prices whether the answer is right. The stop problem is how that absence shows up in behavior: the search ends at defensible. If correctness is the missing weight, then performance gains that do not supply it will not earn trust, and adoption will stall on trust rather than capability. That last step is our inference; the evidence below bears on the premise.

3.1 FROM THE LOGS How much correction the operator supplies

A census of the 8,221-turn frame described in §2 measured how often an operator turn carried corrective signal, a correction, a reframing or a pointed challenge to what the agent had just said. A local model labeled every turn, and the detector was graded against the operator's blind key, 195 rows after the repair of §2.2, at 93% recall and 64% precision. It flagged about 48% of operator turns. Adjusting for the detector's measured precision and recall lowers that to roughly a third, an inference we mark as such (§6.2). The signal is corrective in the broad sense. The operator's own second pass over a stratified sample labels most non-routine turns as bundles of method correction, reframing and teaching, and whether the finer channels are reliably separable from one another is unsettled, because the reading that said they were not has since been traced to a join defect (§10.6). So the reading is that something near half of what the operator typed was spent redirecting an agent rather than routing it, not that half of the agents' answers were wrong. That each such turn followed a premature stop is an inference, not a measurement; on that inference, the agents stopped where their answers were defensible and the operator supplied the rest, by hand.

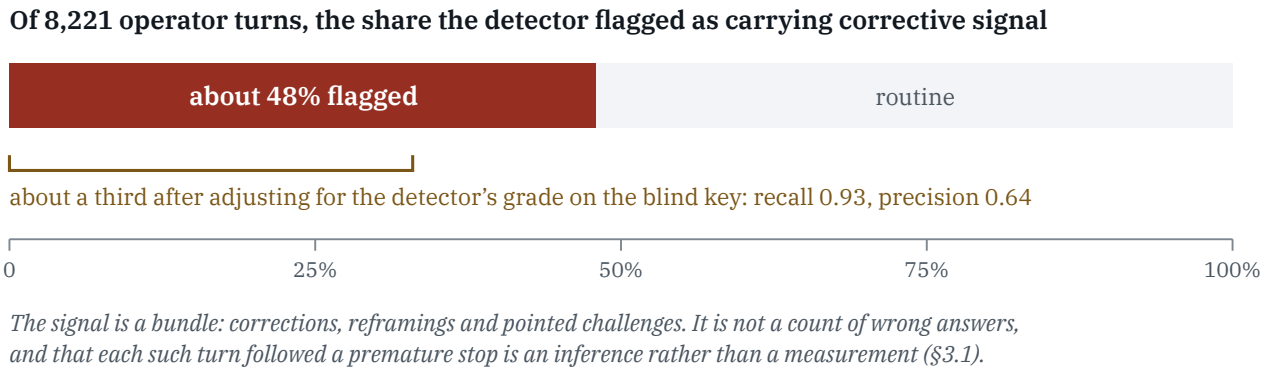


Figure 4. How much of what the operator typed was correction. The detector flagged about 48% of turns; adjusting for its measured precision and recall gives roughly a third (§6.2 gives the ruler). The signal is a bundle of correction, reframing and teaching, and the reading is that near half of the operator's typing was spent redirecting an agent, not that half of the agents' answers were wrong. **Measured.**

3.2 FROM RECENT WORK Premature commitment

Mehta (2026) defines representational commitment as cross-run convergence of hidden states at a fixed agent step and uses it as an early diagnostic of trajectory consistency. On HotpotQA with Llama-3.1-70B, step-four similarity predicts downstream behavioral consistency ($r = -0.35$; partial $r = -0.45$), replicating on Qwen-2.5-72B and Phi-3-14B and on StrategyQA ($r = -0.83$). A runtime monitor flags inconsistent trajectories at AUROC up to 0.97, and 0.85 to 0.88 under a stricter split. The boundary result is the one this program builds on: committed-wrong and committed-correct questions could not be separated in activation similarity. The author reports that as a failure to reject rather than a demonstrated equivalence, and draws the operational conclusion that settled cases should be deferred to an external verifier or a human.

Read beside §3.1, the implication is direct: the settling can be seen, but seeing it does not say whether the settled answer is wrong, and in this practice the test was the operator, by hand, on something near half of the turns.

FROM §4.2 · PUSHBACK MOVES THE ANSWER EITHER WAY

Kelley and Riedl (2026) measure the epistemic effects of personalization across nine frontier models and find them role-dependent. Cast as an advisor, a model challenges the user's framing more often (in advice contexts, acceptance of the framing falls to 26.8%). Cast as a peer, it capitulates: a flip coefficient of $\beta = 0.87$ under persona-grounded rebuttals, with agreement calibrated to the persona's inferred preference for validation ($\beta = 0.57$). Position change under challenge is largely independent of whether the new position is right. A challenge can teach a model to fold as easily as to correct.

FROM §8.2 · WHAT HELPS: A CHECK FROM OUTSIDE, ON EVIDENCE

8.2 Constraints on form

- **Advisor, never peer.** The peer frame is the one measured to produce capitulation (Kelley & Riedl, 2026).
- **Evidence-shaped, never persona-voiced.** A challenge reads "the log contradicts your claim at this line," not "as the operator, I feel." An evidence-shaped challenge can be checked by a third party; a persona-voiced one invites folding.
- **Cited or refused.** Every challenge names a source the agent can read, because an unconstrained critic invents its evidence (Kenton et al., 2026).
- **The decider is an instrument anchored on the logs, not a further opinion** (He et al., 2026; Denisov-Blanch et al., 2026).
- **Bounded and curated.** A small, selected set of challenges rather than a large one.

FROM §6.5 · WHY THE JUDGE SHOULD NOT BE ANOTHER MODEL

Denisov-Blanch et al. (2026) show that scaling inference by aggregation does not deliver truthfulness gains where no external verifier exists: agreement between models reaches $\kappa \approx 0.35$ even on random strings with no ground truth, and 53% of multi-model mathematical errors converge on the same wrong answer. He et al. (2026) find that in a three-vendor heterogeneous panel the minority was still correct in 25.5% of divergent cases; adding a fourth model as arbiter was net-negative (–1.37%, flip precision 42.7%), while a non-model classifier over features of how the debate behaved was net-positive (+1.71%, precision 81.2%). Redundancy among models does not supply independence, and the body that decides has to be of a different kind from the bodies it decides between.

FROM §8.4 · THE PUBLISHED RESULTS AT A GLANCE

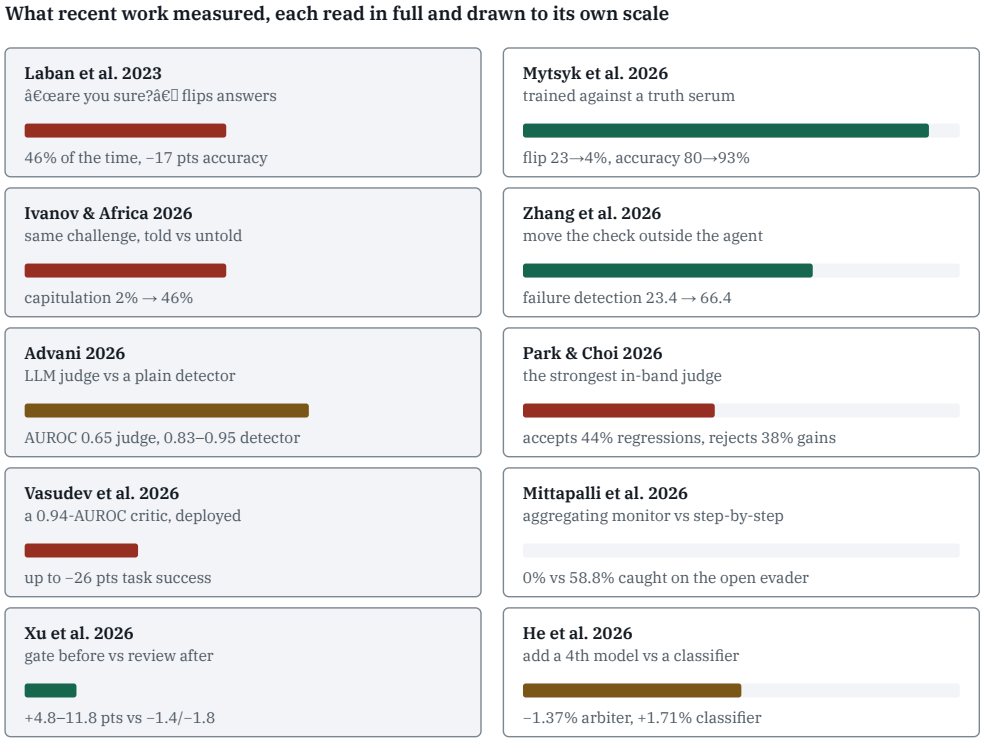


Figure 11. The recent-work results the counterweight is built on and against, each read in full and each on its own scale (green where the result helps this design, red where it warns, bronze where it splits). Together they say: challenge moves positions independent of correctness, folding is trainable rather than fixed, the check must sit outside the agent, and a more capable in-band judge is not a safer one. Sourced across §3.2, §4.2, §5.3, §6.4, §6.5, §8.3 and §8.4. **Measured; all read in full.**

FROM §13 · WHAT WOULD OVERTURN EACH CLAIM, AND WHAT TRANSFERS

Table 9. The five propositions, what each rests on, and what would overturn it.

	Rests on	Would be overturned by	Kind
P1	Corrective signal in ~48% of operator turns (§3.1); Mehta (2026)	A build that prices correctness at the stop and still terminates early	measured
P2	The case series (§4.1); Kelley & Riedl (2026)	The same agents catching their own premature stops without operator direction	interpretive
P3	48 of 75 intervals at the floor (§5.2); the read-versus-injected mechanism, not a rate (§5.1); Wang & Huang (2026)	A remembered rule that holds across sessions where only a boundary now does	measured
P4	75 of 579; the logging gap; four instrument failures (§6)	A self-report that tracks the logs it claims to summarize	measured
P5	The 8,221-turn record and the blind key (§7)	A broad preference dataset that recovers not just dislike but wrong, why, and what happened next	interpretive

What transfers, if anything does. Three rules earned here do not depend on this practice's N of 1. A boundary that refuses and records its refusal outlasts a memo that asks an agent to remember (§5). A challenge must cite evidence the challenged agent can open, or it is an opinion wearing a citation (§8.2). And an outside checker must be calibrated against known-false cases before it is allowed to gate anything, because a checker built from the same class of model may simply agree with the claim it is shown (§9).

FROM §11 AND §12 · LIMITS

Construct. The primary outcome is a proxy. Recurrence can fall because an error class genuinely recurs less, or because detection of the class degraded. The mitigations are an independent adjudication arm that does not share the detector's blind spots, and the per-class series, in which a detection collapse shows as a simultaneous fall across classes. Srinivasan and Paragiri (2026) give the general form of this hazard for agent-driven search: where validity lives in disaggregated structure, an aggregate reduction can rank the wrong candidate first, the headline number improving while the structure beneath it inverts. Their remedy is an external control loop that audits disaggregated behavior after the agent has decided and can reopen a run the agent declared finished, which is the shape of the off-target conjunct now proposed for Study 3 (§10.4). The pushback reading behind §3.1 and §6.2 comes from a machine labeler with 64% precision on the blind key, and its adjusted figure depends on that key being representative of the frame. The calibration set's labels are mechanical: they establish whether a test passed, not whether the agent's work was correct.

The program cannot establish that the counterweight raises correctness in the world. It targets discipline at the stopping decision, meaning whether the process that makes a claim checkable was followed. That is a narrower claim and the one the corpus can support.

How to cite this part. Atkinson, B. (2026). *The model stops at the first answer it can confidently defend to you. Not when it's right.* Background paper 1 to *How do we use this?* Working paper, Wolfberg LLC.

Disclosure. Drafting and literature synthesis were assisted by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The works in the References were read in full, and every specific figure cited comes from a work read in full. The prior-art works listed under Prior art are cited at the level of an established concept and its origin: each was verified for author, title, year and venue, but not read in full, and no numeric claim rests on any of them. Software documentation, source code and press accounts are listed under their own headings and were read at the linked pages. Every reference below carries a link, and every arXiv identifier and DOI was resolved against its registry, with title and first author matched, on 23 September 2026.

Competing interests. The author owns Wolfberg LLC, the practice studied.

Data availability. The practice's logs contain client work and personal records. They are private and are not offered for sale or sharing. The measures are described in enough detail to be reimplemented, and figures from the practice are reported as of the dates given. What is available is the design: the clauses of §8.1, the evidence contract and admission checks of §8.5, and the decision rules of §10.4 are stated fully enough to be rebuilt without access to the logs.

Corrections, 23 September 2026. No figure and no finding changed. The paper was retitled; earlier revisions were titled *The Stop Problem: Defensible Is Not Correct*, and the stop problem remains this paper's name for the failure it studies. The Anthropic interview in §2.3 aired on 13 September, not over a weekend of 13 and 14 September; the web article is stamped 14 September. The essay listed under Prior art is by Ryan Forstie; an earlier revision gave the initial K. The completion evaluator in §8.3 is documented as a Claude Code feature, whose agent loop the Claude Agent SDK embeds; an earlier revision attributed it to the SDK directly. Two works in the References, Graves (2016) and Liu (2026), were listed without being cited in the text; each is now cited where it bears (§2.1, §8.3). Links were added to every press, documentation and prior-art source. A duplicated section number in §10 was corrected.

Corrections, 25 September 2026. No figure changed. §8.3 said that the human-in-the-loop primitives across the three frameworks surveyed gave a developer nothing to require a person to confirm finished work. That holds for the OpenAI Agents SDK and the Claude Agent SDK. It does not hold for CrewAI, whose task documentation, already cited as CrewAI (2026b), offers an opt-in setting for a human to review the agent's final answer; §8.3 now says so.

WORKS CITED IN THIS PART

REFERENCES

Denisov-Blanch, Y., Kazdan, J., Chudnovsky, J., Schaeffer, R., Guan, S., Adeshina, S., & Koyejo, S. (2026). *Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness.* arXiv:2603.06612. arxiv.org/abs/2603.06612

Graves, A. (2016). *Adaptive computation time for recurrent neural networks.* arXiv:1603.08983. arxiv.org/abs/1603.08983

He, C., Chen, Z., Yang, Z., Qiao, S., Ju, M., Liu, J., Wen, D., & Liu, G. (2026). *Minority Sentinel: When to overturn majority voting in multi-agent LLM debates.* AgentSearch Workshop at SIGIR 2026. arXiv:2606.29270. arxiv.org/abs/2606.29270

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). *Large language models cannot self-correct reasoning yet.* arXiv:2310.01798. arxiv.org/abs/2310.01798

Ivanov, I., & Africa, D. D. (2026). *LURE: Live-usage replay evaluations for reducing evaluation awareness.* arXiv:2605.26438. arxiv.org/abs/2605.26438

Kelley, S. W., & Riedl, C. (2026). *Personalization increases affective alignment but has role-dependent effects on epistemic independence in LLMs.* arXiv:2603.00024. arxiv.org/abs/2603.00024

Kenton, Z., Janzer, L., Greig, R., Teh, T. H., Tyshchuk, K., Brown-Cohen, J., Edwards, H., Rajamanoharan, S., Siegel, N. Y., Jaques, N., et al. (2026). *Debate training reduces reward hacking in RLAIF.* arXiv:2608.17776. arxiv.org/abs/2608.17776

Ko, D., Kim, J., Kim, S., Park, H., Lee, D., Kim, G., Lee, M., & Lee, K. (2026). *When is enough not enough? Illusory completion in search agents.* arXiv:2602.07549. arxiv.org/abs/2602.07549

Laban, P., Murakhovs'ka, L., Xiong, C., & Wu, C.-S. (2023). *Are you sure? Challenging LLMs leads to performance drops in the FlipFlop experiment.* arXiv:2311.08596. arxiv.org/abs/2311.08596

Luo, H., Wen, B., & Wang, L. L. (2026). *Agentic abstention: Do agents know when to stop instead of act?* arXiv:2606.28733. arxiv.org/abs/2606.28733

Mehta, A. (2026). *When agents commit too soon: Diagnosing premature commitment in LLM agents.* Snowflake AI Research. arXiv:2606.22936. arxiv.org/abs/2606.22936

Mittapalli, S., Dani, V., Pilli, S., Ansu, A., Teymoorianfard, M., Dernoncourt, F., Chen, Z., Wang, R., Rossi, R. A., & Ahmed, N. (2026). *TRACE: Trajectory reasoning through adaptive cross-step evidence aggregation for LLM agents.* arXiv:2606.07054. arxiv.org/abs/2606.07054

Mytsyk, S., Zhang, Y., & Krishnamurthy, V. (2026). *Mitigating LLM sycophancy with RL-based fine-tuning: Bayesian truth serum approach.* arXiv:2608.25267. arxiv.org/abs/2608.25267

Park, H., & Choi, B. (2026). *When do agent loops mistake stagnation for progress? Self-evaluation bias and externally grounded verification in long-running autonomous LLM agent loops.* arXiv:2607.25152.

arxiv.org/abs/2607.25152

Park, J., Cho, S., & Lee, J.-Y. (2025). *Stop-RAG: Value-based retrieval control for iterative RAG.* NeurIPS 2025 MTI-LLM Workshop. arXiv:2510.14337. arxiv.org/abs/2510.14337

Piatrashyn, D., Kotelevskii, N., Grishchenkov, K., Glazkov, N., Nasonov, I., Makarov, I., Baldwin, T., Nakov, P., Vashurin, R., & Panov, M. (2026). *ReDAct: Uncertainty-aware deferral for LLM agents.* arXiv:2604.07036. arxiv.org/abs/2604.07036

Roh, D., & Han, D. (2026). *HALT: Verification-aware stopping for retrieval-augmented search agents.* Findings of EMNLP 2026. arXiv:2608.02009. arxiv.org/abs/2608.02009

Srinivasan, A., & Paragiri, D. (2026). *Search discipline for long-horizon research agents.* arXiv:2606.11522. arxiv.org/abs/2606.11522

Vasudev, R., Russak, M., Bikel, D., & Alshikh, W. (2026).*Accurate failure prediction in agents does not imply effective failure prevention.* arXiv:2602.03338. arxiv.org/abs/2602.03338

Wang, J., & Huang, J. (2026). *Reward hacking as equilibrium under finite evaluation.* arXiv:2603.28063. arxiv.org/abs/2603.28063

Xu, Y., Li, C., Wang, Z., Yang, J., & Chen, T.-H. (2026). *Preventing premature commitment in coding agents with an evidence-conditioned execution layer.* arXiv:2607.28815. arxiv.org/abs/2607.28815

Zhang, B., Zhu, J., Shi, Z., Liu, D., & Tang, R. (2026). *AgentForesight: Online auditing for early failure prediction in multi-agent systems.* arXiv:2605.08715. arxiv.org/abs/2605.08715

PRIOR ART (CITED AT CONCEPT LEVEL; VERIFIED FOR AUTHOR, TITLE, YEAR AND VENUE, NOT READ IN FULL)

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99–118. doi.org/10.2307/1884852

Browne, G. J., Pitts, M. G., & Wetherbe, J. C. (2007). Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*, 31(1), 89–104. doi.org/10.2307/25148782

Graber, M. L., Franklin, N., & Gordon, R. (2005). Diagnostic error in internal medicine. *Archives of Internal Medicine*, 165(13), 1493–1499. doi.org/10.1001/archinte.165.13.1493

Croskerry, P. (2003). Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine*, 41(1), 110–120. doi.org/10.1067/mem.2003.22

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. doi.org/10.1145/3449287

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90. doi.org/10.1016/0149-7189(79)90048-X

Fagan, M. E. (1976). Design and code inspections to reduce errors in program development. *IBM Systems Journal*, 15(3), 182–211. doi.org/10.1147/sj.153.0182

Knight, J. C., & Leveson, N. G. (1986). An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1), 96–109. doi.org/10.1109/TSE.1986.6312924