

The Self-Verification Gap: Why Working Hallucination Detectors Are External, and What Real-Time Correction Would Actually Require

Pranay M. Mahendrakar

2026-08-20

Abstract

Large language models generate fluent text that is sometimes false, and a substantial literature now proposes to have the model notice and repair those errors while it writes. Parts of the problem are settled. Semantic entropy detects confabulations across datasets and tasks without prior knowledge of the task (Farquhar et al., 2024). Search-augmented verification agrees with crowdsourced annotators 72% of the time on roughly 16,000 individual facts, at more than 20 times lower cost (Wei et al., 2024). Probes on hidden states separate true from false statements at 71-83% balanced accuracy in distribution (Azaria and Mitchell, 2023). The tension is that every one of these signals is external to anything the model reports about itself, and that the four properties a deployed detector needs - low cost, transfer under distribution shift, span-level localisation, and a false-positive rate low enough to act on - have never been demonstrated together in a single system. This paper states five constraints such a system must satisfy, argues that no published system satisfies more than two, and proposes five experiments that would settle the open questions. We report no experiments and claim no result of our own; nothing here establishes that such a system can be built.

1. Introduction

The case against real-time self-correction is easy to overstate, so begin with what already works. Semantic entropy - an entropy computed over meaning-equivalence clusters of sampled generations rather than over token sequences - detects confabulations across datasets and tasks without prior knowledge of the task (Farquhar et al., 2024, in Nature). Search-augmented verification of decomposed atomic facts agrees with crowdsourced human annotators 72% of the time on approximately 16,000 individual facts, at more than 20 times lower cost (Wei et al., 2024). Probes trained on hidden activations classify statements as true or false at 71-83% balanced accuracy within their training distribution (Azaria and Mitchell, 2023, preprint). Truthfulness information is not diffuse: it concentrates in specific tokens of the generated answer (Orgad et al., 2024). Hallucination detection has working instruments.

What those instruments have in common is the problem. Semantic entropy is a statistic over multiple samples, not a report the model makes about itself; search-augmented verification consults an external index; a probe is a separately trained classifier reading activations the model does not itself consult. Every component here with demonstrated traction sits outside the introspective loop, and the strongest direct test of introspection concerns a model predicting its own behaviour rather than assessing its own claims (Binder et al., 2024, preprint). The working parts of hallucination detection are external. That is not a claim that introspection is worthless, and Section 2 takes care not to make one.

The deeper difficulty is that the demonstrated properties do not co-occur. Semantic entropy transfers across tasks but carries a 5- to 10-fold increase in inference cost (Kossen et al., 2024, preprint), and its authors scope it to confabulations rather than to hallucinations generally. The cheap signals are the representation probes, and that family performs close to random out of distribution, with in-distribution performance on RAGTruth attributed in part to a spurious correlation with the data (Dubanowska et al.,

2025); their own limitations section declines to call such transfer impossible, so the correct summary is that it is undemonstrated for this family rather than ruled out. Localisation is where the numbers collapse: on the RAGTruth Overall split, prompted GPT-4-turbo reaches 63.4 F1 at response level but 28.3 F1 at span level with 18.4% precision (Niu et al., 2024). Self-assessment carries an authorship bias, with confidence inflated by up to 26% on the model's own answers relative to identical user-supplied answers (Sanz-Guerrero et al., 2026). Not cheap, not transferable, not localised, not introspective: no published system demonstrates more than two of these four together.

A detector also implies a gate, and a gate implies an operating point. The field reports AUROC, which hides it: failure-prediction AUROC of 0.522 for black-box verbalized confidence against 0.605 white-box appears as a routine comparison rather than as a failure case (Xiong et al., 2024). The cost of firing wrongly is meanwhile measurable - prompting a model to edit its own biographies drops FActScore from 66.7 to 58.6 (Mishra et al., 2024). No paper in the corpus we surveyed reports false-positive rate at usable recall, at span granularity, under distribution shift, at a fixed generation-time compute budget. That is a gap, not a disagreement.

The open problem, then, is not whether models can self-correct. It is a specification question:

Can a per-claim trustworthiness signal be produced that simultaneously satisfies five constraints - (i) computable inside a generation-time compute budget, (ii) retaining discriminative power under distribution shift, (iii) localised finely enough to support a targeted repair, (iv) with a false-positive rate low enough that acting on it is net-positive given the cost of a wrong intervention, and (v) rendered to a downstream consumer in a form that improves decision quality rather than merely acceptance - and if not, which constraint should be relaxed, and at what measured cost?

The contributions are four. First, Section 3 states what the negative results establish and what they do not: the central refutation concerns unaided, prompted, whole-response correction on closed-form reasoning tasks (Huang et al., 2024), and long-form factual generation is a different setting on which the evidence is thinner. Second, Section 4 decomposes the pipeline into seven functions - generate, estimate, localise, verify, correct, express, and decide - records for each whether the evidence is intrinsic or external, and closes by setting the resulting configurations against the five constraints; the seventh function is largely absent from the systems literature. Third, Section 5 sets out the error-composition and operating-point questions the field has not asked. Fourth, Section 6 proposes five experiments, each stated so a reader could run it, with Section 7 giving the methodological preconditions any such evaluation must meet.

We present no experiments and report no measurements of our own; every number here is attributed to the work that produced it. The claim is that the problem is specifiable and that affordable experiments would settle it, not that the system we specify would work. Section 8 states what would falsify the agenda, including the possibility raised by Kalai et al. (2025, preprint) that the right intervention is to change how models are scored rather than to add machinery to them.

2. Scope and terminology

Four terms carry the argument of this paper, and each is used here more narrowly than in common usage.

Hallucination denotes a generated claim that is not supported by whatever the generation is accountable to: the supplied context in a grounded setting, the world in an open-domain setting. *Confabulation* is a proper subset. Farquhar et al. (2024) scope semantic entropy to "a subset of hallucinations - confabulations - which are arbitrary and incorrect generations", meaning errors that vary across resamplings of the same prompt. The distinction is not cosmetic: every consistency-based detector, from

SelfCheckGPT (Manakul et al., 2023) to semantic entropy, is defined on the variable class. Blindness to an error the model repeats stably follows from that premise; it is an inference we draw, not a claim those authors make. This paper addresses the broader hallucination category and treats the split between the two classes as an unmeasured quantity rather than a settled one.

Self-signal we define as any per-response or per-span reliability estimate read off the generating model itself. Two kinds must be kept apart, because the literature routinely conflates them. The first is a trained, externally-grounded self-signal: Asai et al. (2023, preprint) attach reflection tokens whose supervision comes from outside the model, and Kossen et al. (2024) train probes on hidden states whose regression target is semantic entropy, itself computed from multiple samples and an entailment model. In both cases the "self" names where the signal is read, not how it was learned. The second is an introspective self-report: the model is asked, in context, to judge its own output using nothing beyond its own forward pass. The distinction is the pivot of this paper.

Binder et al. (2024) is the paper that tests introspection most directly, and it deserves its due. They report a real privileged-access effect - a model fine-tuned to predict properties of its own behaviour outperforms another model trained on the same behavioural data. But the effect fails on more complex tasks and on out-of-distribution generalization, and, decisively for our purposes, it concerns self-prediction of behaviour, not assessment of whether a claim is true. It is not evidence that a model can verify its own factual output.

Kadavath et al. (2022, preprint) exhibit the pattern that recurs throughout the rest of this paper. They report encouraging performance, calibration and scaling for $P(\text{True})$, a self-evaluation signal - and they note that it improves when the model conditions on many of its own sampled answers. The lift arrives with cross-sample structure. We suspect this is the general shape of the field's positive introspection results: something external or aggregate supplies the discriminative content, and the model supplies the interface.

This is why the working title of this project, which used the word "self-aware", was dropped. The term is not a neutral label for a capability. It presumes an introspective mechanism, and the mechanism is what the evidence does not support.

3. What the negative results establish, exactly

Three papers carry most of the pessimism in this area, and each is narrower than its reputation.

Huang et al. (2024) define intrinsic self-correction as the model revising its initial response "based solely on its inherent capabilities, without the crutch of external feedback," and measure it on GSM8K, CommonSenseQA and HotpotQA. GPT-4 on GSM8K goes 95.5 -> 91.5 -> 89.0 across two correction rounds; GPT-3.5 on CommonSenseQA goes 75.8 -> 38.1 -> 41.8. Two cells are flat rather than down at round one and one rebounds slightly at round two, so the exact statement is that no benchmark ends above its baseline, not that every cell decreases monotonically. The paper also attributes earlier positive results to contamination: oracle labels in RCI and Reflexion, an unfair comparison against self-consistency in multi-agent debate, sub-optimal prompt design in Self-Refine. Supply the oracle label and correction does help (GPT-3.5 GSM8K 75.9 -> 84.3), which suggests the deficit sits in knowing whether to revise rather than in revising.

Six things that result does not establish. First, task form: all three benchmarks have a single extractable discrete answer, and open-ended long-form factual generation - the actual target of a hallucination corrector - is untested. Second, externally grounded correction is explicitly out of scope, and where tested it improves on the unaided baseline. Gou et al. (2023) report that tool-interactive critiquing consistently improves performance; Valmeekam et al. (2023, preprint) raise Blocksworld plan accuracy from 55% with an LLM self-verifier to 88% with the sound verifier VAL; Stechly et al. (2025) lift Game-

of-24 from 5% to 36% and Graph Coloring from 16% to 38% with a sound verifier, on tasks where self-critique drives the same models to 3% and 2%. Third, trained correction: Kumar et al. (2024) instil self-correction through multi-turn online reinforcement learning on self-generated traces, improving base-model self-correction by 15.6% on MATH and 9.1% on HumanEval, and Kamoi et al. (2024) reach the same conclusion in survey form. Fourth, granularity: whole-response regeneration is not span-local editing. Mishra et al. (2024) test the latter on factual text, and the intrinsic negative reproduces at span level while a retrieval-augmented trained editor improves the text instead; Section 4.4 gives that comparison in full. It is the citation for a claim about factual generation, because Huang et al. tested reasoning. Fifth, impossibility: the title ends in "Yet." Sixth, Huang et al. do not separate detection from repair.

Tyen et al. (2024) do, and the asymmetry is load-bearing for everything that follows. Their models "generally struggle" to find logical mistakes "even in highly objective, unambiguous cases," yet correction is robust across five reasoning tasks once the error location is supplied, and a small classifier trained on out-of-domain data outperforms a prompted large model at mistake-finding. This indicates that the binding constraint is localisation rather than repair. Their correction result is obtained with ground-truth error locations injected, so it bounds what a perfect detector would buy, not what an achievable one will.

Dubanowska et al. (2025) is the detection-side analogue, and its scope is equally specific: linear probes on model internals, a random forest classifier and sparse-autoencoder probes, auditing ReDeEP and SAPLMA. Out-of-distribution generalisation "is currently out of reach, with all of the analyzed methods performing close to random," and state-of-the-art performance on RAGTruth "is largely driven by a spurious correlation with data." Their own Limitations section must travel with that sentence: out-of-distribution generalisation "is currently not achieved by any method we are aware of but the introduction of new detection methods might change this state of affairs, so it is not a definite statement on the impossibility of such methods." A live tension runs against it in any case: Farquhar et al. (2024) report that semantic entropy generalises robustly to tasks not seen before. Different method family, different protocol, so the two are not straightforwardly contradictory. The defensible claim is that out-of-distribution transfer is undemonstrated for representation probes, not that it is unattainable.

Three adjacent results are routinely quoted past their warrant. Jiang, D., et al. (2025) show that models are not reliably better at discriminating among previously generated alternatives than at generating initial responses - the absence of a reliable advantage, not evidence that discrimination is worse. Kalai and Vempala (2024) prove a lower bound for arbitrary facts unresolvable from training data, for models satisfying a generative calibration condition, at a rate close to the Good-Turing fraction of facts appearing exactly once; they find no such pressure for facts appearing more than once or for systematic facts such as arithmetic, and write that such models "may require post-training to mitigate." Ma et al. (2026) demonstrate a gradient conflict between accuracy and calibration objectives under reinforcement learning from verifiable rewards; that is a demonstrated conflict, not an impossibility theorem.

The counter-tension deserves equal weight, because the negative literature is itself narrow. It rests on GPT-3.5 and GPT-4-era models, three or four benchmarks each, one or two correction rounds, and closed-form tasks with decidable ground truth. Almost none tests long-form open-ended factual generation. Three recent results suggest that part of the measured deficit may be an activation or formatting artefact rather than a capability ceiling: Tsui (2025) measures a 64.5% self-correction blind spot across 14 open-source non-reasoning models, reduced 89.3% by appending "Wait"; Chen, K.-Y., et al. (2026, preprint) raise explicit error-flagging rates by 23 to 93 percentage points, significant in 10 of 12 settings, purely by relabelling the message role of a byte-identical claim; Sanz-Guerrero et al. (2026) improve calibration by up to 26% by reframing the model's own answer as user input, with no

retraining. All three use injected or attributed errors, on non-frontier or non-reasoning models, and one is unrefereed. We present this as a hypothesis with a named test (Section 6.3), not as an established reframing.

What has been refuted is the strongest version of the hypothesis: unaided, prompted, whole-response, single-model self-correction on closed-form reasoning. What remains open is almost everything a deployed system would do - externally grounded, span-local, trained, provenance-controlled correction of long-form factual text. Silence is not promise, and the burden is on the proposer to name which dimension carries the residual and what evidence would show it.

4. Component analysis

A framework that promises to "detect and correct hallucinations" names an outcome, not a mechanism. We decompose that outcome into six functions - generate, estimate trustworthiness, localise the suspect span, verify, correct, express confidence - and argue in Section 4.6 that the evidence forces a seventh, decide. For each we ask what the evidence indicates is achievable today, what the best measured result is and in whose numbers, and whether the function can be discharged by the generating model reasoning over its own output or needs a signal produced outside it. The third question is where this paper's thesis lives, and we state a verdict on it for every component.

4.0 Generate: the base rate every later stage must beat

Detection is worth its cost only relative to how often the generator is wrong. ChatGPT reaches 58% atomic factual precision on people biographies under FActScore (Min et al., 2023). In RAGTruth, 43.1% of responses contain at least one hallucinated span - 14,289 spans across 17,790 responses (Niu et al., 2024). In deployed generative search engines, only 51.5% of sentences are fully supported by their citations (Liu, N. F., et al., 2023); on ELI5 in ALCE, the best systems lack complete citation support half the time (Gao, T., et al., 2023). Below all of these sits the Good-Turing floor of Kalai and Vempala (2024), which applies to the singleton-fact class only and carries the exemptions set out in Section 3.

Verdict: not a detection component. *It fixes the base rate, and at 58% atomic precision a corrector worth a few points does not produce a trustworthy system.*

4.1 Estimate trustworthiness at the response level

Asking the model how confident it is produces a signal that is real but too weak to gate on. Xiong et al. (2024) report near-chance failure prediction from black-box verbalized confidence - the 0.522 AUROC quoted in Section 1 - and report it as a routine comparison, not a failure case. Zhou et al. (2024) find an average of 47% of the responses a deployed language model marks as confident to be erroneous. Mahaut et al. (2024) find confidence often unstable across semantically equivalent inputs, and Yona et al. (2024) add a faithfulness gap distinct from miscalibration. The estimate is contaminated by authorship as well, at the magnitude Section 3 reports (Sanz-Guerrero et al., 2026).

The opposite evidence carries equal weight. Kadavath et al. (2022) report encouraging performance, calibration and scaling for $P(\text{True})$, improving when the model conditions on many of its own samples, in the pattern set out in Section 2. Tian et al. (2023) find verbalized confidence better calibrated than the models' own conditional probabilities for RLHF models including ChatGPT, GPT-4 and Claude, often reducing expected calibration error by a relative 50%; Yoon et al. (2025) find reasoning models strictly better calibrated in 33 of 36 settings. The defensible claim is that introspective confidence is insufficient, authorship-biased and overconfident in absolute terms, not that it is uninformative.

Alignment also appears to damage the substrate the estimate is read from. Figure 8 of the GPT-4 Technical Report - a figure caption, not an abstract-level claim - shows MMLU expected calibration

error moving from 0.007 pre-trained to 0.074 after PPO, with the caption stating that post-training hurts calibration significantly (OpenAI, 2023).

External estimators do better in distribution and worse out of it. Hidden-state probes label true and false statements at the 71-83% balanced accuracy quoted in Section 1 (Azaria and Mitchell, 2023), and Mahaut et al. (2024) rank trained probes the most reliable family, at the cost of access to weights and training data. Semantic entropy transfers across tasks at the 5- to 10-fold computation cost Kossen et al. (2024) describe, and representation-based detectors are close to random out of distribution on the audit reported in Section 3 (Dubanowska et al., 2025).

Verdict: not feasible intrinsically. *Feasible today only as a separately trained, externally calibrated module - and even then near-random out of distribution for the probe family.*

4.2 Localise the suspect span

This is the component that breaks, and the number showing it break is the most important one this paper quotes. The RAGTruth Overall figures quoted in Section 1 decompose as follows: prompted GPT-4-turbo scores 63.4 F1 at the response level - precision 46.9, recall 97.9 - against 28.3 F1 at the span level with 18.4% precision (Niu et al., 2024). Two readings follow. First, the response-level score is bought by flagging nearly everything: at 97.9% recall the detector approximates a constant "yes", and a system that answers "yes" is not a gate. Second, at the granularity a repair operator needs, better than four in five fired flags point at text that is not hallucinated. Prompted GPT-3.5-turbo falls to 12.8 span F1; a purpose-finetuned Llama-2-13B reaches 78.7 response-level and 52.7 span-level F1 on the same split, a large gain that still leaves span-level F1 near 50.

The collapse reproduces across benchmarks built by different groups. On FELM at segment level, GPT-4 scores 35.4 F1 vanilla, rising to 48.3 with reference documents, while ChatGPT scores 4.9 F1 at 49.7 balanced accuracy, which is chance; chain-of-thought is not monotonically helpful there, lowering ChatGPT to 3.8 (Chen, S., et al., 2023). On FavaBench, FAVA reaches 48.1 average fine-grained F1 against GPT-4 at 34.2 and ChatGPT at 18.8 (Mishra et al., 2024). On FaithBench - adversarially filtered to cases where detectors already disagreed, so its figures are not general detector performance - the best F1-macro is 54.90, from MiniCheck-Deberta-L, against GPT-4o at 40.75 (Bao et al., 2025). Small purpose-trained verifiers are better here, not merely competitive. These numbers are also not comparable across papers as a matter of course: Janiak et al. (2025, preprint) report established methods dropping by up to 45.9% when scored under a human-aligned LLM-as-Judge protocol rather than by lexical overlap.

The ground truth is itself unstable. Mu-SHROOM reports inter-annotator IoU of 0.486 for English and 0.874 for Italian, best systems at 0.65 and 0.78 (Vazquez et al., 2025); the organisers attribute part of that spread to larger annotator pools, since the statistic falls mechanically with pool size, so 0.486 is protocol-dependent and not a clean human ceiling. One result from the same task is decisive for architecture rather than measurement: the XLM-R token-classification baseline, mapping tokens to hallucination probabilities without any intermediate fact-verification step, ranks at or near the bottom of every language leaderboard, below the trivial heuristic of marking everything. That is precisely a naive learned self-signal head in the sense of Section 2, and it is the one design of that kind on which a direct measurement exists.

Verdict: external and trained. *The best span-level results reported here are roughly 50 F1 with a dedicated model and roughly 28 F1 at 18.4% precision when prompted, both on RAGTruth Overall, and token classification without an external verification step is specifically demonstrated insufficient.*

4.3 Verify

Checking a claim against evidence has the strongest empirical base of any component here, and none of the working implementations are introspective.

SAFE, which splits a long-form response into individual facts and issues search queries for each one, is the system behind the annotator-agreement and cost figures quoted in Section 1; it also wins 76% of a randomly drawn 100-case disagreement subset (Wei et al., 2024). MiniCheck-FT5, at 770M parameters, reaches GPT-4 accuracy on grounded fact-checking, and the method delivers GPT-4-level performance for 400 times lower cost (Tang et al., 2024). That is the single most important measurement for any real-time claim, and it comes with a hard boundary: MiniCheck verifies against a grounding document that is already supplied, and does not retrieve. Retrieval, which SAFE performs once per atomic fact, is where the latency sits.

The verifier is also a large error term in its own right. On AttrScore, zero-shot GPT-4 reaches 85.1 overall micro-F1 at judging whether a cited reference supports a statement; ChatGPT reaches 55.0, and fine-tuned models span 47.4 to 75.2 (Yue et al., 2023). A single accuracy number is not a safety specification, because verifier errors are asymmetric and the direction of the asymmetry flips between domains. GPT-4 self-verifying Blocksworld plans is 61% accurate with a true-positive rate of 98.2% and a false-positive rate of 84.45%, accepting 38 of 45 genuinely invalid plans (Valmeekam et al., 2023). Stechly et al. (2025) report the mirror-image failure: 72.4% accuracy on Graph Coloring at a 95.8% false-negative rate, and 79.6% on Mystery Blocksworld at 97.09%.

Two constraints bind the evidence channel. A second call to the same model is not an independent judge: self-preference correlates with self-recognition capability (Panickssery et al., 2024), and self-refinement improves fluency while amplifying self-bias (Xu et al., 2024). And evidence supply is non-monotone - top-ranked non-relevant documents hurt while random documents improved accuracy by up to 35% (Cuconasu et al., 2024), and mid-context evidence is used unreliably even by long-context models (Liu, N. F., et al., 2024).

***Verdict: external, and feasible now.** The component with the strongest empirical base, external by construction: the evidence comes from outside the model, and the best cost-accuracy point is a 770M-parameter verifier that is not the generator. The open variables are verifier independence and retrieval latency, and no operating point is reported for any of these verifiers under distribution shift.*

4.4 Correct

Asking the model to rewrite its own output is the operator with the most direct evidence against it, and the evidence bearing on factual generation - as opposed to reasoning - is the editing comparison in Mishra et al. (2024). Prompting ChatGPT to edit ChatGPT's own biographies drops FActScore from 66.7 to 58.6. Llama2-Chat as editor drops it to 52.6, and the same editor takes Alpaca-13B output from 42.5 to 22.7. On the ChatGPT-generated biographies the retrieval-augmented trained editor is the only condition that improves the text, reaching 70.0; on the Alpaca-13B row the prompted editors improve it only once retrieval is added. Self-Refine's own table is the fairest evidence for task-dependence: its math rows move 64.1 to 64.1, 74.8 to 75.0, and 92.9 to 93.1, while the headline average is carried by Dialogue Response and Sentiment Reversal, tasks with no hard correctness criterion (Madaan et al., 2023).

Correction anchored to an external signal behaves differently. RARR researches claims against web search and post-edits unsupported content, significantly improving attribution while otherwise preserving the input (Gao, L., et al., 2023). Reflexion, anchored to executable feedback, reaches 91% pass@1 on HumanEval against a prior GPT-4 state of the art of 80% (Shinn et al., 2023). Substituting a sound verifier for the model's own plan critique produces the Blocksworld gain reported in Section 3 (Valmeekam et al., 2023), and training instills the behaviour directly (Kumar et al., 2024).

We argue the more promising move is to change the operator rather than improve the rewriter. Conformal back-off progressively reduces claim specificity, offering 80 to 90% correctness guarantees while retaining the majority of the original output (Mohri and Hashimoto, 2024, preprint); conformal abstention supplies theoretical guarantees on the hallucination rate (Abbasi Yadkori et al., 2024, preprint). Two cautions attach. Any correction loop must be scored against a compute-matched baseline, and in the closed-form settings where that comparison has been run it is unflattering (Section 5.2). And the robustness of repair given a supplied error location (Section 3) bounds what a perfect detector would buy, not what an achievable one will.

***Verdict: external signal required, harmful without one.** Correction is feasible given correct localisation and grounded evidence, and measurably damaging when driven by the model's unaided judgement of its own text. Regeneration is probably the wrong operator; specificity reduction and abstention are better matched to the available signal.*

4.5 Express confidence

The elicitation literature does not support a simple ordering of channels. The near-chance verbalized-confidence AUROC in Section 4.1 sits against the opposite direction in Tian et al. (2023), where for models fine-tuned with human feedback verbalized confidence is better calibrated than the models' own conditional probabilities. Neither result licenses a claim that one channel dominates.

The human-facing evidence does establish that the default rendering is counterproductive: explanations increase the chance people accept a recommendation regardless of its correctness (Bansal et al., 2021), and longer explanations raise user confidence even when the extra length does not improve accuracy (Steyvers et al., 2025). Neither hedging language nor numbers repair this on their own. First-person hedges reduced agreement and improved accuracy while semantically similar impersonal phrasing produced weaker, non-significant effects, and overreliance was reduced but not eliminated (Kim et al., 2024); numerical certainty aided error detection, but reliance fell far behind the stated certainty, and combining numbers with verbal explanation amplified blind trust rather than reducing it (Papenkordt et al., 2025). The benefit is conditional: explanations reduce overreliance when they lower the cost of verifying a specific claim (Vasconcelos et al., 2023).

***Verdict: external, and a design problem rather than a modelling problem.** The channel has to be specified against a stated objective and user-tested. Fluent prose plus a number is measurably the wrong default.*

4.6 Decide and abstain

Calibration and action are dissociated, which is why deciding has to be a separate component rather than a consequence of the previous one. Across 20 frontier models and 20 datasets, AbstentionBench finds abstention an unsolved problem on which scaling is of little use, and finds reasoning fine-tuning degrading abstention by 24% on average, including in the math and science domains those models are explicitly trained on (Kirichenko et al., 2025). Wang, J., et al. (2026, preprint) find that even where extreme penalties make abstention mathematically optimal, models almost never abstain, and utility collapses.

***Verdict: external policy, not emergent behaviour.** The least-studied component and possibly the binding one. Nothing in the evidence suggests it can be left to the model.*

4.7 The pattern across components

Laid side by side, the components sort cleanly. Verify has a demonstrated cost-accuracy point and is external by construction; estimate, localise, correct, express and decide have none that is intrinsic.

Every component with measured traction moves work outside the generator - to retrieval, to a separately trained small verifier, to a supervised probe, to a conformal wrapper, or to an explicitly specified policy - and at the two places where the intrinsic route has been tested at the granularity a repair loop needs, span-level localisation (Section 4.2) and span-level editing (Section 4.4), prompted detection flags roughly four wrong spans for every right one and unaided editing lowers FActScore. The pattern is not that introspection carries no information. It is that no introspective signal has yet been shown to support an action.

Set against the five constraints of Section 1, no configuration in this literature clears more than two at once. MiniCheck-FT5 clears the compute constraint and, given a supplied grounding document, operates at claim granularity - but it does not retrieve, and no operating point is reported for it under shift. Semantic entropy clears transfer, within the confabulation class, but not compute: the cost is 5 to 10 times inference. The purpose-finetuned span detector of Section 4.2 comes closest on localisation, but is trained and evaluated in distribution, with no precision reported at a fixed compute budget. We are aware of no published system that clears three. None has been evaluated on the fifth constraint at all: the rendering evidence in Section 4.5 comes from a separate human-factors literature that does not use these detectors.

5. Error composition and the operating-point problem

A detect-localise-verify-correct loop is not one system with one error rate. It is a chain, and the literature reports numbers for the links but not for the chain. Two things follow: nobody can say where a pipeline's residual error comes from, and nobody can say when it should act on what it detects.

5.1 The six-stage stack

Take the minimal pipeline. A generator produces text, a decomposer splits it into checkable units, a decontextualizer makes each unit self-contained, a retriever supplies evidence, a verifier judges support, and a corrector edits. Every stage carries its own error rate, and each was measured on different data by a different paper.

The generator's base rate is the best characterised of the six (Section 4.0). The decomposer is not neutral preprocessing. Wanner et al. (2024) show FActScore is sensitive to the decomposition method, so decomposer error is charged to the generator's account; Hu et al. (2025) report that decomposition's effect on fact-checking is not uniformly positive; and Core (Jiang, Z., et al., 2025) shows sub-claim precision metrics can be manipulated by adding obvious or repetitive sub-claims. The decontextualizer inherits the structural tension Wanner et al. (2025) identify: augmenting a sub-claim so it can be checked in isolation leaves it ambiguous which part of the augmented claim the verifier's judgement covers. The retriever is non-monotone in both relevance and position, and the verifier is a large error term whose errors are asymmetric in domain-dependent directions; Section 4.3 gives both sets of figures. One accuracy number for a verifier is not a safety specification.

No end-to-end error model for this pipeline exists in the corpus we surveyed. Without one, a benchmark gain cannot be attributed to the component its authors credit, and a failure cannot be diagnosed. Writing one - stage error rates as free parameters, a composition rule, and a falsifiable prediction about which stage dominates at a given base error rate - requires no experiments, and we propose it as work the field can start now.

5.2 The operating point

A detector implies a gate, because something must decide when the corrector fires, and a gate requires an operating point: a committed false-positive rate at a chosen recall. AUROC names no such point. It averages over thresholds nobody would deploy and is invariant to the base rate that decides whether

firing is net-positive at all. The failure-prediction AUROCs in Section 1 say that one signal orders errors slightly better than another; they do not say where to cut. The span-level figures in Section 4.2 make the gap concrete, since more than four in five flagged spans there are not hallucinations.

The gap named in Section 1 can now be placed exactly. Each of its four qualifiers - a false-positive rate at usable recall, at span granularity, under distribution shift, at a fixed generation-time compute budget - appears somewhere in this literature. They never appear together.

A false positive is not free, and its cost is measurable in FActScore points: the unconditioned editing runs collected in Section 4.4 lose between 8 and 20 points. Those editors ran without a precision-controlled gate, so the deltas bound the cost of indiscriminate intervention rather than isolating the per-false-positive cost. The cost is nonetheless finite and measured, which is what a break-even analysis needs. Compute enters that analysis jointly with precision rather than as a separate budget line: the two cost-accuracy anchors are semantic entropy and MiniCheck-FT5 (Sections 4.1 and 4.3), and the obvious amortisation - retrieve once, verify many sub-claims against one context - is blocked by Akhter et al. (2026, preprint), who find decomposition helps only under strictly sub-claim-aligned evidence and degrades under repeated claim-level evidence.

Finally, the compute-matched baseline for open-ended generation is undefined. In closed-form settings the comparison exists and is unflattering: multi-agent debate scores 83.0 on GSM8K against self-consistency's 88.2 for the same nine responses (Huang et al., 2024, Table 7), and plain resampling at $k=25$ beats a sound-verifier-plus-critique loop on Game of 24, 42 against 38, and on Graph Coloring, 44 against 38 (Stechly et al., 2025, Table 1). But self-consistency presupposes an extractable discrete answer (Wang, X., et al., 2023), which long-form factual generation does not supply. The nearest anchor is a sampling-based detector such as SelfCheckGPT (Manakul et al., 2023). Defining that baseline is itself an open problem, and until it is defined no reported gain from a hallucination pipeline can be read as a gain over spending the same tokens on something simpler.

6. Five open research questions

Each question below is unanswered in the literature reviewed above and investigable with existing corpora and models. For each: why it is open, a design, and what would settle it either way. We have run none of them.

6.1 RQ1. What fraction of deployed hallucinations manifest as sampling inconsistency?

The consistency-based family rests on a premise about the phenomenon, not about the detector: SelfCheckGPT treats hallucinated content as content on which independent samples diverge and contradict one another (Manakul et al., 2023), and semantic entropy is scoped to the confabulation subset defined in Section 2 (Farquhar et al., 2024). Simhi et al. (2025, preprint) establish that a high-certainty error class exists but do not measure its prevalence. That split is unmeasured, and it upper-bounds the recall of every sampling-based detector, and so the whole cheap-detection branch.

Design. On a span-annotated corpus of natural hallucinations - RAGTruth (Niu et al., 2024), FavaBench (Mishra et al., 2024), FaithBench (Bao et al., 2025) - resample the generator k times under the same prompt, cluster the samples by bidirectional entailment with an off-the-shelf NLI model following Kuhn et al. (2023), and label each annotated span variable or stable. Report prevalence with confidence intervals, stratified by domain and error type.

What settles it. A prevalence figure, plus per-class recall for semantic entropy and SelfCheckGPT. If variable spans dominate, sampling-based detection is a viable core with a documented blind spot. If stable spans dominate, the cheap branch is a niche and a real-time design must be retrieval-first.

6.2 RQ2. What false-positive rate is achievable at usable span-level recall, under shift and at fixed compute, and where does firing become net-harmful?

AUROC hides the operating point a gate has to choose, and the evidence assembled in Sections 4 and 5 brackets that point badly: 18.4% span-level precision for a prompted frontier model, verifier false-positive and false-negative rates that flip direction between domains, near-chance failure-prediction AUROC, and a measured cost of 8 to 20 FActScore points for editing without a gate. Nobody has expressed the net-harm threshold as a function of base error rate, error asymmetry, repair-success probability, and iteration count.

Design, in two halves. Derive the break-even condition analytically, with detector false-positive and false-negative rates, base hallucination rate, and $P(\text{repair succeeds} \mid \text{correct flag})$ as free parameters, and publish the surface as a falsifiable prediction. Then locate real detectors on it - LLM-AggreFact, RAGTruth, FaithBench - under shift control rather than held-out splits (Dubanowska et al., 2025), human-aligned metrics rather than lexical overlap (Janiak et al., 2025, preprint), a length heuristic the benchmark must defeat, and a fixed token and latency budget with a compute-matched comparator (Section 5.2). Report false-positive rate at fixed recall.

What settles it. The curve, plus where deployed detectors sit relative to it: an argument about whether correction helps becomes a threshold a system either clears or does not.

6.3 RQ3. Does provenance stripping survive contact with naturally occurring hallucinations and reasoning-trained models?

The three results in Section 3 - the ownership bias measured by Sanz-Guerrero et al. (2026), the blind spot measured by Tsui (2025), and the role-relabelling effect reported by Chen, K.-Y., et al. (2026, preprint) - indicate that part of the self-correction deficit may be an artefact of how self-attributed content is represented rather than a capability ceiling. All three use injected or attributed errors, on non-reasoning or non-frontier models, and one is unreviewed. We suspect part of the effect is an artefact of the injected-error paradigm; that is conjecture.

Design. A factorial over presentation (own draft shown as an internal thought, a user message, a tool response, or with provenance stripped), error origin (injected discrete error, or natural hallucination sampled from annotated spans), training regime (reasoning-trained or not), and access (open-weight or frontier API). Outcomes: explicit flagging rate and span-level F1.

What settles it. Whether the effect size survives the natural-error, frontier-reasoning cell. If it does, provenance representation is a lever the field has not pulled. If it vanishes, the reframing collapses and the pessimistic reading stands.

6.4 RQ4. Is out-of-distribution transfer failure a property of the representation or of the supervision target?

Dubanowska et al. (2025) point their own path forward at spans rather than general labels. Orgad et al. (2024) find probes failing across datasets while error type remains predictable from internal state. Ying et al. (2026, preprint) find probes transferring across most truth domains, failing on sycophantic and expectation-inverted cases, and recovering under joint multi-domain training. Kossen et al. (2024, preprint) report semantic-entropy probes generalising better out of distribution than probes trained to predict accuracy directly. That comparison is the tell: it indicates the failure may live in the label rather than the representation. No study varies the label while holding everything else fixed.

Design. Hold architecture, base model and training corpus constant and vary only the supervision target - response-level correctness, span-level annotation, error type, semantic-entropy regression, sufficient-context labels (Joren et al., 2024, preprint) - at matched parameter count and inference cost, evaluated

in distribution and on held-out corpora under shift control.

What settles it. An ordering of supervision targets by out-of-distribution retention. If distilled and span-level targets retain while response-level correctness collapses, the remedy is architectural and actionable. If every target collapses together, representational pessimism is confirmed at a strength Dubanowska et al. explicitly decline to claim.

6.5 RQ5. Do calibration, team decision quality, and abstention-adjusted utility select the same system?

Three success criteria are in play, three literatures optimise them separately, and no study has run one system under all three. Vodrahalli et al. (2022) present an initial exploration that suggests displaying an already well-calibrated model as more confident can improve final human-AI performance. Band et al. (2024) ground calibration in downstream user prediction rather than introspection. Bo et al. (2025) show reduced over-reliance and appropriate reliance coming apart. And the abstention results in Section 4.6 show that a system with a usable confidence estimate can still fail to act on it.

Design. Fix one detector and one repair operator, and render its output three ways: an honest per-claim probability; a confidence transform learned from human incorporation data; and a localised span flag with retrievable evidence attached and no prose justification, per the verification-cost principle of Vasconcelos et al. (2023). On the same participants and tasks, measure expected calibration error, two-sided appropriate reliance, and utility under stated asymmetric penalties. Add a longitudinal arm, since Li et al. (2025) report user self-confidence aligning to AI confidence and persisting after the AI is withdrawn.

What settles it. Whether the three objectives rank the three renderings differently. If they do, reporting calibration alone is measuring the wrong thing, and showing that is itself a result.

Two further problems are engineering-grade rather than question-shaped, but they gate any implementation. The first is the retrieval tension: per-sub-claim retrieval is the central latency obstacle, and the amortisation that would remove it is blocked from both sides, by Akhter et al. (2026, preprint) on the decomposition side and by Liu, N. F., et al. (2024) on the long-context side. The second is decomposition as an error and gaming surface, set out in Section 5.1. Neither has a published solution; any proposed system must budget for both.

7. Methodological preconditions for any evaluation of such a framework

Every study proposed above can be run badly enough to produce a number that means nothing. Six requirements follow from results already in the literature; a study violating one of them cannot settle the question it was designed to settle.

Shift-controlled splits, not held-out splits. The spurious correlation Dubanowska et al. (2025) identify in state-of-the-art RAGTruth performance (Section 3) means a held-out split from the same corpus cannot separate a detector from a corpus artefact. The split must be built to break a specific correlation, and that correlation must be named.

Human-aligned metrics, not lexical overlap. Janiak et al. (2025, preprint) report the drop quoted in Section 4.2 when established methods are scored against a human-aligned LLM-as-Judge protocol rather than by lexical overlap. A method whose rank order depends on which of the two is used has not been evaluated.

A trivial-baseline gate. Janiak et al. also report that simple heuristics based on response length can rival complex detection techniques. Any benchmark licensing a claim about a detector must be one on which a length heuristic and a mark-everything heuristic both fail, and both must be published beside

the headline number. Mu-SHROOM supplies the precedent, in the token-classification baseline that finishes below mark-all (Section 4.2). Curated true/false sentence sets of the kind Azaria and Mitchell (2023) use make surface confounds easier to control than long-form output does, and the control is correspondingly harder to claim on long-form output.

Decomposer ablated separately from verifier. Decomposition is a first-class error source, not preprocessing (Section 5.1), so an end-to-end number that does not separate the two assigns error to the wrong stage. Decontextualization is a third such stage (Wanner et al., 2025).

An informativeness filter. Precision-style factuality scores can be raised by adding obvious or repetitive sub-claims; Jiang, Z., et al. (2025), in the Core framework, identify this and propose informative sub-claim selection as the remedy. A pipeline that controls its own decomposition and is then scored on the resulting claims must report the filter it applies, or the metric is gameable.

A fixed compute budget with a compute-matched baseline. Neither of the compute-matched comparisons in Section 5.2 exists unless compute is held fixed, and for open-ended generation the matched baseline is not obvious for the reasons given there. Choosing it is itself an open problem.

Harnesses exist for most of this: LM-Polygraph (Vashurin et al., 2025), LLM-AggreFact (Tang et al., 2024), ALCE (Gao, T., et al., 2023), HELM (Su et al., 2024, preprint; the hallucination-evaluation harness of that name, not the Stanford holistic-evaluation benchmark), RAGTruth (Niu et al., 2024), FavaBench (Mishra et al., 2024), FaithBench (Bao et al., 2025), Mu-SHROOM (Vazquez et al., 2025), LongFact (Wei et al., 2024) and AbstentionBench (Kirichenko et al., 2025). None scores detection under a streaming or latency constraint; all score completed outputs. No public harness makes a detector's score depend on when it fires, and no number quoted in this paper was obtained under a wall-clock budget. Building one is a contribution independent of all five questions above.

8. Threats to this agenda, and what would falsify it

Three internal failure modes and one external one.

(a) If RQ1 finds that confabulations are a minority of deployed hallucinations, the cheap detection branch is a niche. Semantic entropy and SelfCheckGPT would retain their reported performance on that class while missing most of what a deployed system emits. Nobody has measured the split.

(b) If RQ3's provenance effect vanishes on naturally generated hallucinations or on frontier reasoning models, the activation-artefact reframing collapses and the pessimistic reading of the self-correction literature stands unchallenged. We regard this as the most likely of the four: all three results motivating RQ3 use injected or attributed errors.

(c) If RQ4 finds that every supervision target degrades equally out of distribution, no learned trustworthiness head of the family tested is deployable under shift, and the framework has to be rebuilt retrieval-first, at the per-sub-claim latency cost that Akhter et al. (2026, preprint) and Liu, N. F., et al. (2024) jointly impose.

(d) The external threat is Kalai et al. (2025, preprint). If hallucination persists because models are optimized to be good test-takers and guessing when uncertain improves test performance, detection machinery works against the objective the model was trained on, and the effective intervention is rescoring the benchmarks rather than adding architecture. That argument is a position, not a measurement, but nothing we propose answers it.

The portion we claim an architecture could address is the measurement and operating-point one: which signal to compute, at what cost, where to place the gate, what operator to apply when it fires. We concede the incentive portion. If (d) is right, this framework improves a quantity the field is not

rewarding, and the five studies would measure that improvement accurately while it changed nothing.

9. Conclusion

This paper set out five constraints that a real-time hallucination detection-and-repair system must satisfy together, and argued that no published system satisfies more than two of them. Component by component, the parts with measured traction - retrieval, small trained verifiers, supervised probes, conformal wrappers - sit outside the generator's introspective loop, and the least-studied part, abstention, has to be specified as an external policy rather than left to emerge. Two questions the literature has not asked follow from that: how the errors of a six-stage pipeline compose, and where a gate's operating point should sit.

We concede that none of this is settled by argument. Nothing here shows that such a system can be built, and Section 8 concedes that the incentive objection lies outside what any architecture can answer. The five studies in Section 6 are runnable with existing corpora and models, and would settle the measurement and operating-point questions that remain.

References

- Abbasi Yadkori, Y., et al. (2024). Mitigating LLM Hallucinations via Conformal Abstention. arXiv:2405.01563. [PREPRINT]
- Akhter, M. E., et al. (2026). The Alignment Bottleneck in Decomposition-Based Claim Verification. arXiv:2602.10380. [PREPRINT]
- Asai, A., et al. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511. [PREPRINT]
- Azaria, A., & Mitchell, T. (2023). The Internal State of an LLM Knows When It's Lying. arXiv:2304.13734. [PREPRINT]
- Band, N., Li, X., Ma, T., & Hashimoto, T. (2024). Linguistic Calibration of Long-Form Generations. ICML 2024. arXiv:2404.00474.
- Bansal, G., et al. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021, 1-16. doi:10.1145/3411764.3445717; arXiv:2006.14779.
- Bao, F. S., et al. (2025). FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs. NAACL 2025 (Short Papers), 448-461. doi:10.18653/v1/2025.naacl-short.38; arXiv:2410.13210.
- Binder, F. J., et al. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787. [PREPRINT]
- Bo, J. Y., Wan, S., & Anderson, A. (2025). To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. CHI 2025, 1-23. doi:10.1145/3706598.3714097; arXiv:2412.15584.
- Chen, K.-Y., et al. (2026). The Self-Correction Illusion: Role Relabeling Gates Explicit Error Flagging in Large Language Models. arXiv:2606.05976. [PREPRINT - unreviewed]
- Chen, S., et al. (2023). FELM: Benchmarking Factuality Evaluation of Large Language Models. NeurIPS 2023 Datasets and Benchmarks. arXiv:2310.00741.
- Cuconasu, F., et al. (2024). The Power of Noise: Redefining Retrieval for RAG Systems. SIGIR 2024, 719-729. doi:10.1145/3626772.3657834; arXiv:2401.14887.
- Dubanowska, Z., et al. (2025). Representation-based Broad Hallucination Detectors Fail to Generalize Out of Distribution. Findings of the Association for Computational Linguistics: EMNLP 2025, 17563-17575. doi:10.18653/v1/2025.findings-emnlp.952; arXiv:2509.19372. (Last author: Karpowicz, M. P.)
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature 630:625-630. doi:10.1038/s41586-024-07421-0.
- Gao, L., et al. (2023). RARR: Researching and Revising What Language Models Say, Using Language Models. ACL 2023, 16477-16508. doi:10.18653/v1/2023.acl-long.910; arXiv:2210.08726.
- Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. EMNLP

2023, 6465-6488. doi:10.18653/v1/2023.emnlp-main.398; arXiv:2305.14627.

Gou, Z., et al. (2023). CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. ICLR 2024. arXiv:2305.11738.

Hu, Q., Long, Q., & Wang, W. (2025). Decomposition Dilemmas: Does Claim Decomposition Boost or Burden Fact-Checking Performance? NAACL 2025 (Main). arXiv:2411.02400.

Huang, J., et al. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024. arXiv:2310.01798.

Janiak, D., et al. (2025). The Illusion of Progress: Re-evaluating Hallucination Detection in LLMs. arXiv:2508.08285. [PREPRINT - under review]

Jiang, D., et al. (2025). SELF-[IN]CORRECT: LLMs Struggle with Discriminating Self-Generated Responses. AAAI 2025, 24266-24275. doi:10.1609/aaai.v39i23.34603; arXiv:2404.04298.

Jiang, Z., et al. (2025). Core: Robust Factual Precision with Informative Sub-Claim Identification. Findings of the Association for Computational Linguistics: ACL 2025, 19833-19856. doi:10.18653/v1/2025.findings-acl.1018; arXiv:2407.03572.

Joren, H., et al. (2024). Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. arXiv:2411.06037. [PREPRINT]

Kadavath, S., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. [PREPRINT]

Kalai, A. T., & Vempala, S. S. (2024). Calibrated Language Models Must Hallucinate. STOC 2024. arXiv:2311.14648.

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why Language Models Hallucinate. arXiv:2509.04664. [PREPRINT - position paper]

Kamoi, R., et al. (2024). When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Transactions of the Association for Computational Linguistics 12:1417-1440. doi:10.1162/tacl_a_00713; arXiv:2406.01297.

Kim, S. S. Y., et al. (2024). I'm Not Sure, But...: Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. FAccT 2024, 822-835. doi:10.1145/3630106.3658941; arXiv:2405.00623.

Kirichenko, P., Ibrahim, M., Chaudhuri, K., & Bell, S. J. (2025). AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. NeurIPS 2025. arXiv:2506.09038.

Kossen, J., et al. (2024). Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. arXiv:2406.15927. [PREPRINT]

Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. ICLR 2023 (Spotlight). arXiv:2302.09664.

Kumar, A., et al. (2024). Training Language Models to Self-Correct via Reinforcement Learning. ICLR 2025. arXiv:2409.12917.

Li, J., et al. (2025). As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. CHI 2025, 1-16. doi:10.1145/3706598.3713336; arXiv:2501.12868.

Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating Verifiability in Generative Search Engines. Findings of the Association for Computational Linguistics: EMNLP 2023, 7001-7025. doi:10.18653/v1/2023.findings-emnlp.467; arXiv:2304.09848.

Liu, N. F., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics 12:157-173. doi:10.1162/tacl_a_00638; arXiv:2307.03172.

Ma, Z., et al. (2026). Decoupling Reasoning and Confidence: Resurrecting Calibration in Reinforcement Learning from Verifiable Rewards. ICML 2026. arXiv:2603.09117.

Madaan, A., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. NeurIPS 2023. arXiv:2303.17651.

Mahaut, M., et al. (2024). Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. ACL 2024, 4554-4570. doi:10.18653/v1/2024.acl-long.250; arXiv:2406.13415.

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP 2023, 9004-9017. doi:10.18653/v1/2023.emnlp-main.557; arXiv:2303.08896.

Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP 2023, 12076-12100. doi:10.18653/v1/2023.emnlp-main.741; arXiv:2305.14251.

Mishra, A., et al. (2024). Fine-grained Hallucination Detection and Editing for Language Models. COLM 2024. arXiv:2401.06855.

Mohri, C., & Hashimoto, T. (2024). Language Models with Conformal Factuality Guarantees. arXiv:2402.10978. [PREPRINT]

Niu, C., et al. (2024). RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. ACL 2024, 10862-10878. doi:10.18653/v1/2024.acl-long.585; arXiv:2401.00396.

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774. [TECHNICAL REPORT]

Orgad, H., et al. (2024). LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. ICLR 2025. arXiv:2410.02707.

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. NeurIPS 2024. arXiv:2404.13076.

Papenkov, J., Ngonga Ngomo, A.-C., & Thommes, K. (2025). Are numerical or verbal explanations of AI the key to appropriate user reliance and error detection? Behaviour & Information Technology 45:2094-2115. doi:10.1080/0144929X.2025.2568928.

Sanz-Guerrero, M., Mager, M., & von der Wense, K. (2026). Large Language Models Are Overconfident in Their Own Responses. Findings of the Association for Computational Linguistics 2026. arXiv:2606.03437.

Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2023. arXiv:2303.11366.

Simhi, A., et al. (2025). Trust Me, I'm Wrong: LLMs Hallucinate with Certainty Despite Knowing the Answer. arXiv:2502.12964. [PREPRINT]

Stechly, K., Valmeekam, K., & Kambhampati, S. (2025). On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. ICLR 2025. arXiv:2402.08115.

Steyvers, M., et al. (2025). What large language models know and what people think they know. Nature Machine Intelligence 7:221-231. doi:10.1038/s42256-024-00976-7; arXiv:2401.13835.

Su, W., et al. (2024). Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. arXiv:2403.06448. [PREPRINT]

Tang, L., Laban, P., & Durrett, G. (2024). MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. EMNLP 2024, 8818-8847. doi:10.18653/v1/2024.emnlp-main.499; arXiv:2404.10774.

Tian, K., et al. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. EMNLP 2023, 5433-5442. doi:10.18653/v1/2023.emnlp-main.330; arXiv:2305.14975.

Tsui, K. (2025). Self-Correction Bench: Uncovering and Addressing the Self-Correction Blind Spot in Large Language Models. COLM 2026. arXiv:2507.02778.

Tyen, G., et al. (2024). LLMs cannot find reasoning errors, but can correct them given the error location. Findings of the Association for Computational Linguistics: ACL 2024, 13894-13908. doi:10.18653/v1/2024.findings-acl.826; arXiv:2311.08516.

Valmeekam, K., Marquez, M., & Kambhampati, S. (2023). Can Large Language Models Really Improve by Self-critiquing Their Own Plans? arXiv:2310.08118. [PREPRINT]

Vasconcelos, H., et al. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. Proceedings of the ACM on Human-Computer Interaction 7:1-38. doi:10.1145/3579605.

Vashurin, R., et al. (2025). Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. Transactions of the Association for Computational Linguistics 13:220-248. doi:10.1162/tacl_a_00737; arXiv:2406.15627.

Vazquez, R., et al. (2025). SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared Task on Hallucinations and Related Observable Overgeneration Mistakes. SemEval-2025. arXiv:2504.11975.

Vodrahalli, K., Gerstenberg, T., & Zou, J. (2022). Uncalibrated Models Can Improve Human-AI Collaboration. NeurIPS 2022. arXiv:2202.05983.

Wang, J., Zhou, Y., Devic, S., & Fu, D. (2026). Are LLM Decisions Faithful to Verbal Confidence? arXiv:2601.07767. [PREPRINT]

- Wang, X., et al. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023. arXiv:2203.11171.
- Wanner, M., et al. (2024). A Closer Look at Claim Decomposition. *SEM 2024, 153-175. doi:10.18653/v1/2024.starsem-1.13; arXiv:2403.11903.
- Wanner, M., Van Durme, B., & Dredze, M. (2025). DnDScore: Decontextualization and Decomposition for Factuality Verification in Long-Form Text Generation. EMNLP 2025, 23620-23637. doi:10.18653/v1/2025.emnlp-main.1205; arXiv:2412.13175.
- Wei, J., et al. (2024). Long-form factuality in large language models. NeurIPS 37:80756-80827. doi:10.52202/079017-2567; arXiv:2403.18802.
- Xiong, M., et al. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. ICLR 2024. arXiv:2306.13063.
- Xu, W., et al. (2024). Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024, 15474-15492. doi:10.18653/v1/2024.acl-long.826; arXiv:2402.11436.
- Ying, Z. J., Ravfogel, S., Kriegeskorte, N., & Hase, P. (2026). The Truthfulness Spectrum Hypothesis. arXiv:2602.20273. [PREPRINT]
- Yona, G., Aharoni, R., & Geva, M. (2024). Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? EMNLP 2024, 7752-7764. doi:10.18653/v1/2024.emnlp-main.443; arXiv:2405.16908.
- Yoon, D., et al. (2025). Reasoning Models Better Express Their Confidence. NeurIPS 2025. arXiv:2505.14489.
- Yue, X., et al. (2023). Automatic Evaluation of Attribution by Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2023, 4615-4635. doi:10.18653/v1/2023.findings-emnlp.307; arXiv:2305.06311.
- Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. ACL 2024. arXiv:2401.06730.