

Hallucination Across the Reasoning Lifecycle: Interface Visibility, Causal Evidence, and Release Control in Large Reasoning Models

ZHE YU, Zhejiang University, China and Binjiang Institute of Zhejiang University, China

MOHAN LI, Guangzhou University, China

LEI YU, Rensselaer Polytechnic Institute, USA

KA-HO CHOW, The University of Hong Kong, Hong Kong SAR, China

CHENGWEI QIN, The Hong Kong University of Science and Technology (Guangzhou), China

XINGYU WU, The Hong Kong Polytechnic University, Hong Kong SAR, China

WENPENG XING, Zhejiang University, China

SHUGUANG XIONG, Research Center for Scientific Data Hub, Zhejiang Lab, China

MENG HAN, Zhejiang University, China

Reasoning errors can propagate into later decisions and memory. This survey synthesizes 312 papers and first-party reports on text-based reasoning hallucinations around three questions: what evidence is observable, what study designs establish, and which corrective actions the evidence supports. UIPCA records unsupported premises (U), invalid inferences (I), dependent reuse (P), visible answer–trace consistency (C), and action-policy failures (A). Across 58 reviewed sources, no comparison establishes that a specified intervention improves reasoning while reducing factual reliability under matched conditions. The synthesis connects diagnosis to verification, repair, selective release, and persistent-state control across memory, tools, and training feedback.

CCS Concepts: • **Computing methodologies** → **Natural language generation; Machine learning.**

Additional Key Words and Phrases: Hallucination, Large Reasoning Models, Chain-of-Thought, Factuality, Faithfulness, Verification

ACM Reference Format:

Zhe Yu, Mohan Li, Lei Yu, Ka-Ho Chow, Chengwei Qin, Xingyu Wu, Wenpeng Xing, Shuguang Xiong, and Meng Han. Hallucination Across the Reasoning Lifecycle: Interface Visibility, Causal Evidence, and Release Control in Large Reasoning Models. Unpublished manuscript prepared for submission to *ACM Computing Surveys*, 35 pages.

1 Introduction

The same wrong answer can arise from distinct failures. A reasoning model may start from an unsupported premise, make an invalid inference, reuse an earlier error, or contradict its own visible reasoning. Repair may require retrieving evidence, correcting a step, revisiting a dependent branch, checking answer–trace consistency, or abstaining. Final-answer accuracy alone cannot select the appropriate repair.

These errors can also affect later decisions. Agents may store generated content in private or shared memory, act through tools, or retain feedback for later rounds [39, 151, 187]. An unsupported claim can then become future context. Storage alone does not establish propagation: the return path depends on whether the claim is reused before independent repair.

Authors' Contact Information: Zhe Yu, Zhejiang University, Hangzhou, China and Binjiang Institute of Zhejiang University, Hangzhou, China, zyu@zju-if.com; Mohan Li, Guangzhou University, Guangzhou, China, limohan@gzhu.edu.cn; Lei Yu, Rensselaer Polytechnic Institute, Troy, New York, USA, yul9@rpi.edu; Ka-Ho Chow, The University of Hong Kong, Hong Kong SAR, China, kachow@cs.hku.hk; Chengwei Qin, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, chengweiqin@hkust-gz.edu.cn; Xingyu Wu, The Hong Kong Polytechnic University, Hong Kong SAR, China, xingyu.wu@polyu.edu.hk; Wenpeng Xing, Zhejiang University, Hangzhou, China, wpxing@zju.edu.cn; Shuguang Xiong, Research Center for Scientific Data Hub, Zhejiang Lab, Hangzhou, China, xiongshuguang@zhejianglab.org; Meng Han, Zhejiang University, Hangzhou, China, mhan@zju.edu.cn.

A RESEARCH RIVER OF REASONING HALLUCINATIONS

Where support breaks, how error travels, and when a model should abstain

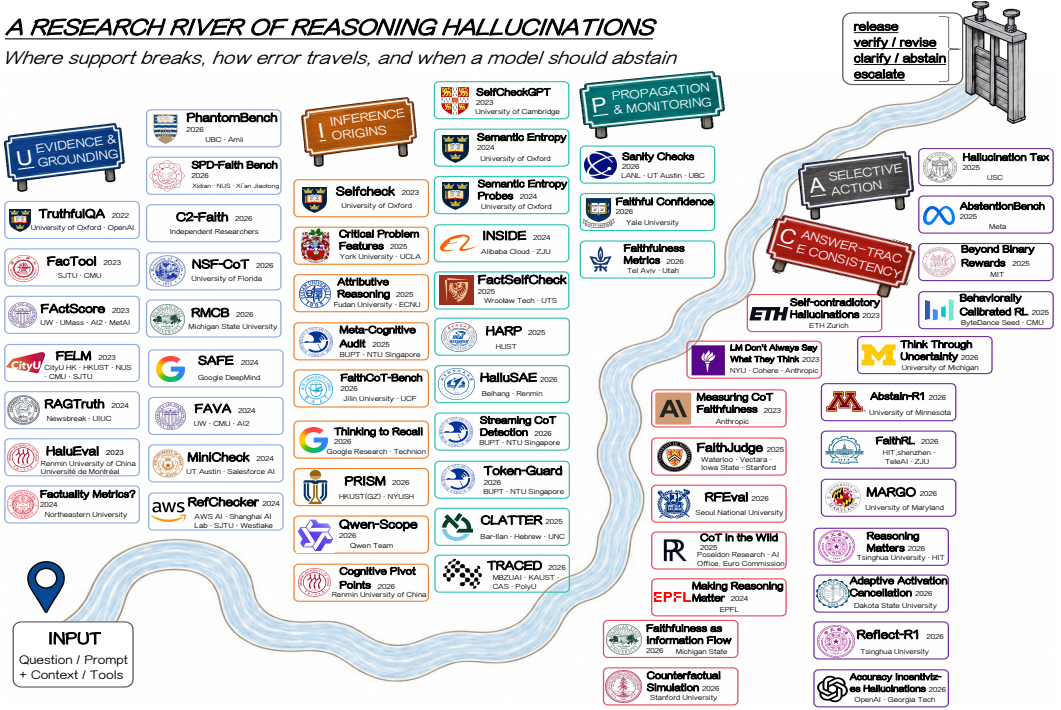


Fig. 1. Research river of reasoning hallucinations: 60 representative works grouped by primary analytical role. The regional headings name research themes, not formal UIPCA fields. In particular, “Propagation & Monitoring” groups related literature; *P* records dependent reuse of a confirmed error, which repetition or a monitoring signal alone cannot establish. Likewise, the consistency region includes process-faithfulness work, while *C* checks only visible answer–trace consistency. The separate action region informs the *A* policy audit. Placement denotes neither chronology nor causal order, and groups are not exclusive. S7 provides canonical titles and links.

Existing taxonomies, monitoring methods, and interventions address different parts of this problem. This survey connects three questions: what the evaluator can observe, what the study design establishes, and which corrective action the evidence supports. Their separation can obscure the difference between consistency and correctness, association and causation, or detection and a justified release decision. Figure 1 maps the literature supporting this evidence-to-action view.

A correct answer can follow unsupported reasoning, while coherent deductions from a distorted graph, mismatched image region, or invented API can yield a wrong answer. Correction creates another distinction: accepting a local edit does not establish recovery of the final outcome [106].

These failures arise in several kinds of reasoning system. Chain-of-thought prompting asks a model to generate intermediate steps; self-consistency samples several reasoning paths and aggregates their answers [82, 169, 173, 201]. Reasoning post-training instead changes the model’s parameters [51, 188]. Whatever the method, a provider may disclose a trace, a summary, or neither [7, 79, 129, 185]. A fluent explanation is therefore something to check, not proof of how the answer was obtained.

These distinctions change how findings should be read. Factual QA can improve or decline across post-training recipes, but such comparisons change complete pipelines [188]. Within a fixed checkpoint, extra deliberation can improve recall even though traces containing a wrong fact have

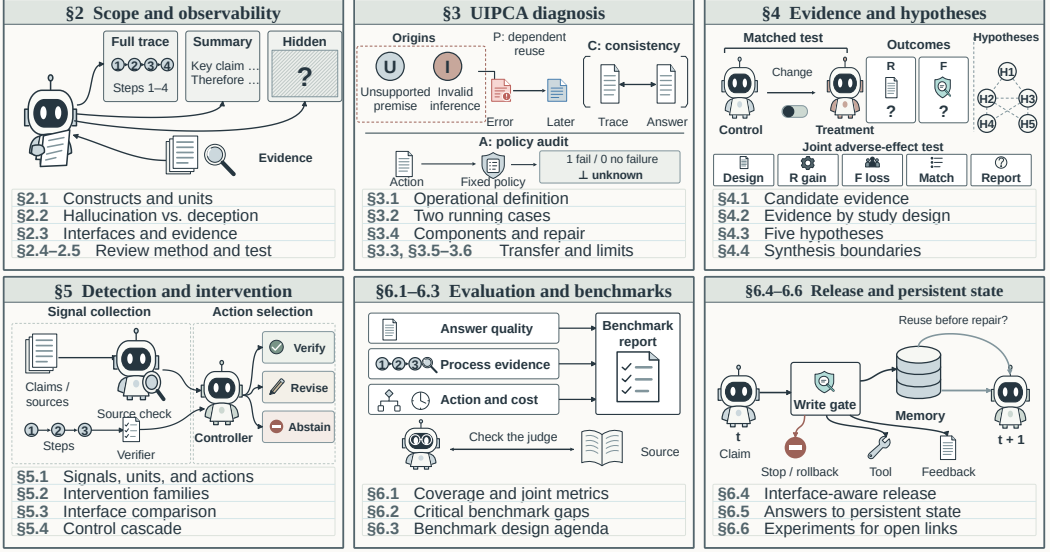


Fig. 2. A guide to the review. Click section indices to open the six topics. In the evidence panel, R denotes reasoning capability and F factual reliability. The five checks concern the joint adverse-effect criterion, not study classes. Dashed links among $H1$ – $H5$ are schematic, not causal claims.

lower final accuracy [45]. The latter association does not isolate the effect of introducing that fact. Neither result establishes that a specified intervention improves reasoning capability while reducing factual reliability.

For consequential claims, the relevant evidence may be a source, image region, formal constraint, or executable test [33, 85, 93, 141]. A final-answer judge cannot recover evidence never exposed, and a detector score alone cannot select an appropriate action. Deployment therefore requires more than detection: it requires an available check and a policy for answering, verifying, revising, or abstaining.

Existing surveys cover adjacent questions but use different organizing units. Broad hallucination reviews organize definitions, causes, benchmarks, detection, and mitigation across NLG and LLMs [64, 68, 197]. General and domain reviews extend that map to foundation models, multimodal failure, and code defects [1, 31, 85, 134, 145]. Reasoning surveys emphasize methods, training, evaluation, and trustworthiness [4, 62, 136, 170].

Specialized reviews organize failures by model lifecycle [57], reasoning type [154], trace length [18], or image-grounded training and task demand [38]. Liu et al. instead derive a taxonomy and test two interventions in an educational application [102]. Table 1 compares their scope and organizing principles.

The review makes three contributions. First, UIPCA provides a common record of error origins, dependent reuse, visible consistency, and policy action. Second, a study-design audit tests the reasoning–reliability evidence against the matched joint adverse-effect criterion. Third, the intervention synthesis compares verification, repair, abstention, escalation, and write control across full-trace, summary, and hidden-reasoning interfaces.

The closest process taxonomy, *Mirages of Logic*, separates premise, operation, logic, and conclusion hallucinations [6]. UIPCA retains such local error types while recording dependent reuse,

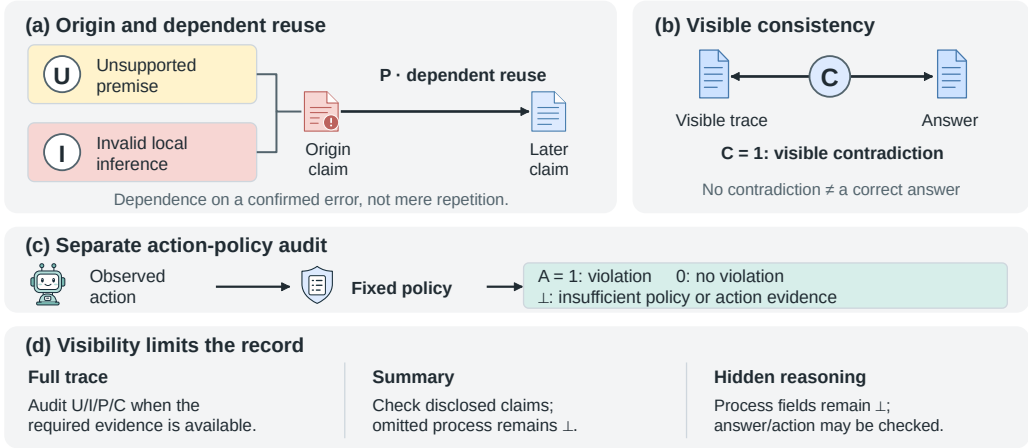


Fig. 3. UIPCA diagnostic record. Panel (a) locates unsupported premises or invalid inferences (U/I) and dependent reuse (P). Panel (b) compares the answer with the visible trace (C), not hidden computation. Panel (c) audits the observed action against a fixed policy (A). Panel (d) shows interface limits: unavailable evidence remains $\perp$, meaning unknown. The fields answer different diagnostic questions, not five sequential stages.

interface-specific unknowns, visible answer–trace consistency, and action-policy failure. These distinctions matter because the same local error may require different repairs depending on its downstream use, observability, and consequences. S4 compares the two schemes on the same cases.

Table 1. What neighboring surveys organize, and which joint gap this review addresses.

Survey / Focus	Interface and Evidence Design	Persistence and Action	What This Review Adds
Broad and domain hallucination surveys [64, 68, 85, 145, 197]	Organize definitions, causes, benchmarks, mitigation, or domain-specific errors; interface disclosure is not the main axis	Controls are usually attached to an output, model stage, or domain failure	Retain their error types and checks, then record visibility, evidence design, dependency reuse, and action
Ho et al. [57]	Multi-level definition and model lifecycle connect text and vision; causal matching is not the main organizing idea	Causes and mitigations span model families, but released claims are not tracked through dependencies or later writes	Add interface-specific unknown labels, study-design classes, and claim reuse
Song et al. [154]	Crosses reasoning types with fundamental, application, and robustness failures	Identifies causes and mitigation directions across informal, formal, and embodied reasoning	Add error location and separate process diagnosis from release-policy failure
Long- and image-grounded CoT surveys [18, 38]	Compare trace length, training method, task demand, or interleaved states	Emphasize capability and method trade-offs rather than later unsupported writes	Use their method distinctions while checking support, visibility, and action
Liu et al. [102]	Evaluates reasoning hallucination in an educational setting and derives a task-specific classification	Tests uncertainty-based abstention and model-to-model correction within that application	Supply the broader, source-level evidence synthesis and interface-aware diagnostic record
<i>Mirages of Logic</i> [6]	Types visible premise, operation, logic, and conclusion failures	Supports local process diagnosis; dependency reuse, interface opacity, and release action are outside its core type labels	Retain error type, then add origin/reuse, unknown-by-interface, answer–trace consistency, and policy action

Figure 3 separates the diagnostic fields, and Figure 8 extends them to retained outputs.

The review asks how to diagnose and repair failures (**RQ1**) and what current evidence establishes about reasoning and factual reliability (**RQ2**). It then examines which detection and intervention strategies different interfaces permit (**RQ3**), and how benchmarks and release policies should evaluate consequential outputs and persistent state (**RQ4**). Sections 2–6 address these questions in turn (Figure 2). S1–S7 provide the corpus, evidence reviews, coding rules, catalogs, and figure data.

Scope. This is a text-centered review of reasoning hallucination. Multimodal and code studies are transfer cases, not co-equal domain surveys. They test which distinctions survive a change in evidence type and where a domain-specific checker is required.

2 Scope, Evidence Model, and Review Method

2.1 Constructs and Evaluation Units

Here, *large reasoning model* means a system that allocates or supervises multi-step computation beyond direct generation. This includes visible scratchpads, hidden effort modes, sampled search, tool use, and reasoning-focused post-training. A displayed rationale is generated text, not necessarily a transcript of hidden computation.

Training methods shape what these systems learn to produce. After pretraining [161], supervised fine-tuning (SFT) learns from demonstrations, while reinforcement learning (RL) rewards selected behavior. Feedback can concern a final answer or individual reasoning steps, and a verifier checks proposed answers or steps [28, 94, 130, 146]. Distillation trains a model using another model’s outputs. In practice, data, rewards, optimization, and the retention of reasoning steps often change together [60, 104, 190, 191]. Labels such as RL-only or SFT+RL describe a training recipe, not a single isolated cause.

In this review, hallucination means generated content that lacks support under the task’s stated reference frame. Two familiar checks capture different aspects of support: *factuality* asks whether a claim is correct, while *source faithfulness* asks whether the supplied evidence supports it [68, 117, 124]. For example, a true statement drawn from outside a designated document may still be unsupported in a document-grounded answer.

Process faithfulness concerns the explanation rather than the source: does the displayed reasoning reflect what actually determined the answer [84]? A coherent explanation can fail this test if the model reached its answer for another reason. Checking the explanation’s statements and testing its causal role are therefore separate tasks.

Confidence and correction add two further questions. A calibrated system’s stated confidence agrees with its observed correctness; corrigibility concerns whether valid feedback repairs later reasoning and the outcome [106]. Neither guarantees a good decision to act. A system may recognize uncertainty yet release an answer that should have been verified or withheld. Table 2 separates these evaluation targets so that success on one is not mistaken for success on all.

2.2 Boundary Between Hallucination and Deception

Hallucination and deception can produce the same false statement, but deception requires evidence of goal-directed concealment or strategic misrepresentation [54, 149]. A false, persistent, or self-contradictory answer does not establish intent. Incentive changes and tests of information access are needed to support a strategic explanation [44, 67].

UIPCA records observable support, dependency, consistency, and action failures without inferring mental state. The same record can arise from missing knowledge, invalid inference, reward pressure, or concealment. A factually supported answer may have no *U/I/P/C* failure and still violate a declared release policy; *A = 1* records that policy failure.

2.3 Observable Interfaces and Evidence Status

The three interfaces differ in what the evaluator can inspect. A full trace exposes ordered reasoning steps; a summary discloses selected claims; a hidden interface exposes no reasoning record. In our comparison, the prompt, designated evidence, answer, and observed action remain available. A summary can assist answer checking, especially when the evaluator lacks the prompt [2], but it cannot establish what happened in omitted steps.

Visibility can also change behavior. Disclosure affects abstention, and length-penalized training can remove monitoring cues with little accuracy loss [98, 120]. Viewing the same run through different interfaces is thus distinct from changing how a model generates or compresses its reasoning.

A missing source has different meanings in different tasks. If an answer must be supported by a supplied document, absence of the required support establishes a grounding failure. In an open-world factual search, failing to find a source does not establish that the claim is false. We therefore distinguish supported, unsupported, and unknown claims. Unknown means that the available evidence cannot decide; it is not a favorable judgment. S4 gives the formal rules.

Table 2. What each evaluation target measures and what the interface permits an evaluator to establish.

Evaluation Target	Operational Question	Typical Unit / Evidence	Interface Boundary
Factuality / source grounding	Is a claim true at evaluation time, or licensed by the designated source?	Atomic claim or answer; fact label, source entailment, execution result	Available under all interfaces when the required external evidence exists
Process faithfulness	Does the displayed reasoning reflect what causally determines the answer?	Trace-answer pairs with controlled perturbations [84, 164]	Causal attribution requires intervention; textual consistency alone is insufficient
Uncertainty calibration	Do reported probabilities or scores match empirical correctness?	Answer set; Brier, ECE, calibration curve [35, 80]	Behavioral confidence can be evaluated under all interfaces; internal signals require provider access
Corrigibility	Does valid feedback change the dependent trace and outcome?	Edited step or trajectory [106]	Requires before/after evidence; local acceptance is not outcome recovery
Selective decision quality	Does a fixed policy choose answer, retrieve, verify, clarify, abstain, or escalate appropriately?	Item and action; selective risk, coverage, action counts	The action is observable; its justification depends on declared policy inputs and evidence

The lifecycle is a separate diagnostic layer, not another evaluation target or model capability. *U/I* locate unsupported premises or invalid inferences, *P/C* describe reuse and consistency, and *A* records a violation of a policy fixed in advance. These fields help interpret the targets in Table 2; they do not replace their measures.

2.4 Search, Selection, and Evidence Strength

This review draws on 312 papers and first-party reports available by 30 August 2026. We combined OpenAlex and arXiv searches with citation tracing and targeted checks of underrepresented topics. Selection focused on reasoning hallucination, its detection and control, and benchmarks that expose relevant failures. Multimodal and code studies test transfer of the framework; they do not form equally comprehensive subreviews.

The collection is curated rather than an exhaustive census of the field. Main-text sources support the synthesis and illustrate consequential differences between methods; the complete catalog retains additional coverage. Figure 4 shows publication years within this collection. Exact queries, eligibility and exclusion rules, screening decisions, and the separate review of reasoning-reliability comparisons are available in S1–S3. The accompanying *Supplementary Methods* PDF explains these procedures without requiring readers to inspect the data files.

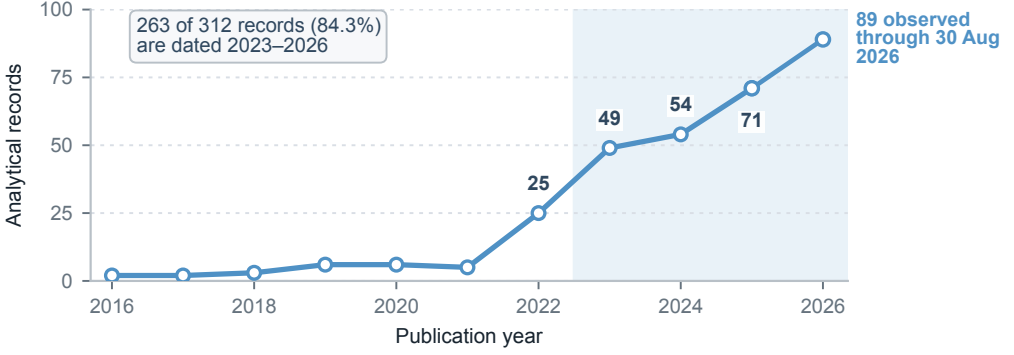


Fig. 4. Publication-year distribution of the 312-record analytical corpus. The solid blue series reaches 89 records observed through 30 August 2026; 263 records (84.3%) are dated 2023–2026. The 2026 count covers only part of the year. These counts describe the curated collection, not exhaustive field-wide publication totals.

For the reasoning–reliability question, relevance screening of all 312 records retained the 20 original S3 sources and identified 14 additional candidates for criterion-level review. Targeted searches and citation tracing from S3 sources sought further relevant comparisons; title-based prioritization yielded 24 additional sources with retrievable primary text for review. These 34 corpus sources and 24 external sources form the criterion-review set, not a second count of the analytical corpus (S3).

Study design determines how these sources inform the argument. Changing one factor in a controlled experiment can test its effect. Comparing complete systems instead reveals a combined outcome, while grouping naturally generated traces shows association. Section 4.2 uses this distinction to explain findings that might otherwise appear contradictory.

Evidence grades distinguish **direct, single-setting evidence** for a stated link, **supported synthesis** across compatible findings, and **limited or mixed evidence** that is sparse or conflicting. A **proposed test** has no result yet; other cases remain **unresolved**.

Use of automated and language-model tools. Deterministic scripts applied the frozen screening rules and generated release ledgers. Language models assisted with sentence-level drafting and prose revision. The authors defined the protocol, selected the analytical sources, verified the evidence used for manuscript claims, determined the analysis, and take responsibility for the submitted text.

2.5 Prospective Matched Joint Adverse-Effect Test

Does improving reasoning come at the expense of factual accuracy? A stronger reasoning model with more factual errors may suggest this trade-off, but the comparison can also reflect different data, rewards, or evaluation tasks. Our *matched joint adverse-effect test* asks a narrower question: does one specified intervention improve reasoning capability while reducing factual reliability?

For comparison j , let ΔR_j and ΔF_j denote intervention-minus-control changes in reasoning capability and factual reliability, both measured on higher-is-better scales. The joint adverse effect is

$$\mathcal{D}_j = \mathbf{1}[\Delta R_j > \delta_R \wedge \Delta F_j < -\delta_F], \quad \delta_R, \delta_F \geq 0. \quad (1)$$

The margins δ_R and δ_F specify the smallest meaningful gain and loss, chosen before evaluation. Equation 1 defines the effect, not its statistical support. To establish it at a prespecified error level α , a joint $1 - \alpha$ confidence region must lie entirely above δ_R and below $-\delta_F$. Multiplicity-adjusted one-sided intervals can enforce the same requirement.

The comparison must isolate the stated treatment: inference-time experiments hold the checkpoint and non-target settings fixed; training experiments start from a common base and control non-target training choices. Both use matched evaluation items and evaluators. Neither trace length nor a “reasoning model” label substitutes for measured capability. Failure to meet the criterion leaves the trade-off unresolved; it does not establish absence of harm. S3 provides implementation and sensitivity rules. We next turn from comparisons between systems to diagnosis within a run.

3 The Evidence-to-Action Lifecycle

UIPCA synthesizes diagnostic distinctions from the reviewed literature into a common record; its five fields do not describe an underlying model mechanism.

UIPCA answers a repair question: once a run is suspect, which observable check can still improve the decision? The *U* and *I* fields locate unsupported premises and invalid local inferences; *P* records dependent reuse. The *C* field compares the answer with the visible trace, while *A* audits the observed action against a fixed policy. The fields therefore serve different diagnostic purposes. UIPCA is neither a flat error taxonomy nor a claim that every run passes through five observable stages. Factual, logical, arithmetic, perceptual, attributional, and executable errors remain separate type labels.

3.1 Operational Definition

Consider a calculation based on a retrieved value. If the value is unsupported, the repair starts with its source (*U*); if the value is correct but the arithmetic is wrong, the repair starts with the inference (*I*). When later calculations reuse the error, correcting only its first occurrence leaves dependent results to repair (*P*). An error mentioned and then discarded does not create this propagation path.

The final answer adds two different checks. Does it agree with the visible reasoning (*C*)? Was releasing it allowed by the policy fixed for the task (*A*)? A wrong answer may agree with a wrong calculation, so consistency alone cannot establish correctness. Conversely, even a correct answer can violate a requirement for independent verification before release. These distinctions connect each diagnosis to a different repair.

In the figures and records, 1 means that the named failure is confirmed, 0 that it is ruled out, and $\perp$ that evidence is insufficient. The first-error marker τ^* locates the earliest confirmed premise or inference failure. It need not locate the actual first error if earlier checks are unresolved. S4 contains the complete coding manual, aggregation rules, and boundary cases.

Visibility limits this record. With hidden reasoning, an evaluator can still check the answer against external evidence and inspect recorded tool actions, but cannot locate errors in unexposed steps. A summary supports checks on what it discloses, not reconstruction of what it omits. Missing process evidence remains unknown, rather than being inferred from confidence or agreement.

UIPCA is a diagnostic record, not an automatic detector. Human judgments, executable checks, and learned detectors can supply its evidence; Section 5 compares their strengths and errors. The running cases below show what the record adds to answer scoring, while Section 3.6 explains its sensitivity to analytical choices.

3.2 Two Running Cases

In Case A, the trace assigns 29 instead of 1 to $1024 \bmod 33$, introduces three more wrong residues, and reuses them in later products [148]. The initial arithmetic error is an *I* failure; later dependent

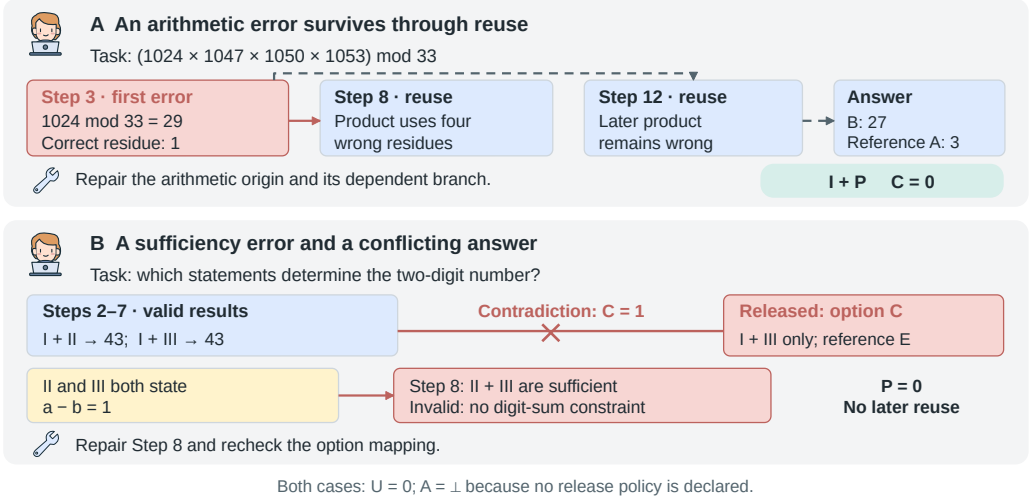


Fig. 5. Two held-out FINE-CoT cases coded with U/I/P/C/A [148]. Case A reuses an invalid Step 3 reduction (I + P); its answer follows the wrong trace, so C = 0. Dashed arrows abbreviate paths through omitted steps. Case B contains a separate Step 8 sufficiency error and a released option that conflicts with valid earlier results (I + C). The crossed line marks this contradiction, not dependence on Step 8. With no later reuse, P = 0.

calculations exhibit P. The answer agrees with the wrong trace, so the consistency check finds no contradiction (C = 0). Correcting the final number alone would leave the faulty derivation intact.

Case B requires a different repair. Steps 2–7 show that I+II and I+III each determine 43. Step 8 wrongly treats II+III as sufficient even though both state only $a - b = 1$ and give no digit sum [148]. This is an inference error without later reuse. Separately, the released option C contradicts the valid earlier results, giving a consistency failure (C = 1). Reference option E establishes that the answer is also wrong. The option comparison matters independently of Step 8.

Answer-only scoring marks both cases wrong but cannot select the repair. Case A needs correction before dependent reuse; Case B needs a constraint and option-mapping check despite no reuse of its Step 8 error. These are worked diagnoses, not estimates of how often either failure occurs. S4 preserves the source traces and reproducible coding checks.

3.3 Text-Centered Evidence and Cross-Domain Transfer Tests

Selected multimodal and code studies test whether diagnostic distinctions survive a change of evidence; they do not estimate domain-wide prevalence or method rankings.

The checks differ: text requires sources, visual claims require objects or regions, and code permits execution without proving every requirement (Table 3).

Table 3. Shared lifecycle with domain-specific evidence and release consequences.

Setting	Premise Object	Early Check	Later Check / Action	Unresolved Boundary
Source-grounded text	Passage, database record, or time-indexed fact	Claim extraction and source entailment [25, 61]	Verify, retrieve, revise, abstain	Hidden evidence use and open-world completeness
Image or video	Object, region binding, frame, or temporal segment	Coverage and grounding [48, 91, 109, 110]	Reasoning check, targeted retrieval, abstention	Whether the rationale causally used the grounded evidence

Table 3. (continued)

Setting	Premise Object	Early Check	Later Check / Action	Unresolved Boundary
Function or repository code	Requirement, API, dependency, program state, repository snapshot	Type/contract/static check [41, 160]	Sandbox, properties, hidden tests, patch review [71, 100]	Untested paths, security, and unwritten intent

Multimodal reasoning makes the evidence boundary explicit. Caption and visual-QA benchmarks detect object hallucinations or image-answer disagreement [13, 73, 144], but do not by themselves distinguish a missed object from an ignored one. Visual-anchor and SPD-Faith analyses examine this distinction [109, 138]. Region-binding and temporal-retrieval methods supply or check visual evidence for subsequent inferences [19, 110]. Medical and chemistry audits add a further warning: a correct answer does not make its displayed rationale valid evidence [89, 182]. Text-only deliberation cannot validate a missed visual detail without renewed access to the image, and answer scoring alone cannot identify the evidence used.

Executable code provides stronger local checks, but no test is universal. Execution exposes mapping, naming, resource, and logic errors. Program-state tasks isolate input-output reasoning, while containers catch invented symbols or imports [41, 47, 160]. Span-level detection across code, tool output, and documents shows why the evidence object must be recorded [205]. Even “tests pass” is conditional evidence. EvalPlus rejects programs accepted by weaker suites and can change rankings [100]. Repository tasks also depend on issue interpretation, file selection, build state, and patch interaction [71]. A code audit must preserve the environment and record any requirements that lack executable tests.

3.4 Lifecycle Components and Their Repair Points

Premise entry (*U*). A premise may be absent, stale, unsupported, or retrieved incorrectly. Short-answer benchmarks show the outcome [74, 172]; cloze probes, knowledge-neuron analyses, retrieval comparisons, and bounded-corpus audits test whether needed information is recoverable [34, 106, 113, 135]. These checks can motivate retrieval or source correction without revealing everything learned in pretraining. Image and code analogues include wrong regions and invented APIs.

Local inference (*I*). Proof audits and program-state tasks reveal invalid transitions despite correct premises [47, 50]. But graph-coloring errors may originate in invented edges, followed by coherent deductions from the wrong graph [56]. The distinction selects input repair versus rule correction. Structured methods separate premise selection and inference [32]; hybrids add symbolic solvers or rule checks [132, 137, 181].

Propagation (*P*). Later claims must depend on a confirmed error; length or repetition alone is insufficient [106, 188]. Independent evidence can stop reuse, whereas self-checks can repeat a shared false premise [114, 123, 126]. In multi-round clinical evaluation, locally coherent rationales sometimes omitted decisive evidence requests before commitment [14]. Repair should therefore examine both acquisition and reuse.

Answer-trace consistency (*C*). A trace can influence an answer or rationalize it, as cue, mediation, perturbation, and decomposition studies show [8, 26, 29, 133, 139, 163]. Behavioral detection is easier for correct than wrong answers [5]; causal-alignment tests instead intervene on trace dependence [171]. The *C* field covers only visible contradiction, not causal attribution or hidden computation.

Action (*A*). Abstention and knowledge-boundary benchmarks reveal unjustified answers and refusals [3, 75, 81]. Reflection training and act-or-abstain tests connect explanations with decisions

[66, 120], but verbal confidence alone is not a policy [76, 95, 118, 180]. Audit false answers, unnecessary refusals, missed clarifications, and escalations against their costs and evidence requirements.

3.5 Applying the Lifecycle Without Overclaiming

A displayed dependency (P) or contradiction (C) need not describe hidden computation. Causal attribution requires changing a premise or step and observing the answer [84, 164]. Textual diagnosis remains useful without treating rationales as mechanism transcripts.

The useful unit depends on the repair. A sentence may be appropriate for checking a source, yet contain several arithmetic operations that need separate checks. A token-level alarm can be precise without identifying which dependent branch should be restarted. Finer localization is useful only when it changes what can be repaired.

U/I/P/C/A records failure locations and relations, not severity. An unused unsupported premise can be serious, while a long failed branch can be harmless if discarded before release. Consequence and repair difficulty therefore remain separate from the number of recorded failures.

3.6 Operational Sensitivity: Segmentation, Dependencies, and Unknowns

A diagnostic record depends on how the trace is divided and which claims are treated as prerequisites. Merging a correct operation with a faulty one broadens the reported first-error location. Treating every neighboring step as dependent can also turn an isolated error into apparent propagation. These are changes in the analysis, not necessarily changes in the model's behavior.

Missing information creates a further limit. A trace can omit a prerequisite or present several derivations for one claim. In such cases, not finding a dependency is not enough to rule out reuse. The affected field remains unknown when the available evidence cannot decide. This is why a complete negative check and an unobserved check have different meanings.

Two laboratory members independently annotated 48 full-trace cases, blind to one another's labels and upstream evaluation labels. With 334 shared source steps, pre-adjudication Cohen's κ was 0.653 for U , 0.730 for I , and 0.858 for P . Case-level C agreement was complete ($\kappa = 1.000$; four jointly positive cases). First-error location agreed in 41/48 cases; requiring the same origin type as well gave 40/48, including joint no-origin judgments. All A labels were unknown because no action policy was supplied. These estimates describe label agreement with fixed segmentation, not validation of the candidate dependency graph. S4 reports case-clustered intervals, positive/negative agreement, disagreement categories, and the status of edge review and adjudication.

RQ1 synthesis. UIPCA helps distinguish a source to recheck, an inference to repair, dependent reasoning to revisit, and a decision to reconsider. Which of these diagnoses is available depends on the interface. An unknown field identifies a limit of the record, not a failure-free part of the run.

4 What the Evidence Actually Establishes

4.1 Candidates for a Matched Joint Adverse Effect

Our criterion-level review covers 34 sources identified within the corpus and 24 additional sources retrieved through external follow-up. No reviewed source establishes the matched joint adverse effect specified in Section 2.5 and Equation 1. Relaxing only threshold and uncertainty-reporting requirements also yields no positive source. These are source-level judgments, not counts of independent experiments. S3 supplies the selection and sensitivity records; Table 4 explains the different conclusions supported by candidate findings.

Table 4. Representative S3 candidates. *R*: separately defined reasoning capability; *F*: factual reliability. Study design and unmet joint-effect criteria are reported separately.

Comparison	<i>R</i> Outcome	<i>F</i> Outcome	Study Design	Unmet Criterion
Post-training recipes [188]	No independent <i>R</i> threshold	Factual QA rises or falls by recipe	Comparison of 12 separately trained systems	Data, reward, and optimizer vary together
System comparisons [20]	System type, not matched <i>R</i>	Higher hallucination for two reasoning systems	Reasoning versus non-reasoning systems	Checkpoint and output detail differ
Natural trace groups [45]	Recall expansion reported separately	Lower accuracy with an incorrect extracted fact	Observed groups on the same questions	Facts are grouped, not experimentally manipulated
Qwen3 thinking mode [168]	No independent <i>R</i>	Aggregate gain with some correct-to-incorrect flips	Fixed checkpoint and items; mode varies	Prompts/decoding differ; beneficial flips coexist
Test-time budget [198]	Tokens, not independent capability	Non-monotonic accuracy and hallucination	Model/task fixed in many budget comparisons	Joint effect unresolved; answer willingness mediates outcomes
Factuality-aware RL [88]	Math reasoning accuracy	Factuality/hallucination measured separately	Training-pipeline comparison	Reasoning and factuality training vary together
Length-penalized CoT [98]	No-hint accuracy, not independent <i>R</i>	Monitorability/faithfulness decline; <i>F</i> unresolved	Length penalty changes the visible trace	Monitorability and <i>F</i> are different outcomes
RFC local edits [106]	No capability outcome	Accepted edits need not change the final outcome	Within-trajectory intervention	Local acceptance and final reversal use different aggregates

The candidates concern different claims: pipeline differences, observed reuse, item-level harm alongside aggregate gains, or monitorability rather than factuality. Table 4 identifies the missing link in each comparison; absence of a strict positive is a design-and-reporting finding, not a field-wide null result.

4.2 Evidence by Design, Not by Headline

Pipeline-confounded comparisons. Among 12 paired systems, all four cold-start SFT+RL pairs improve on both factual QA sets; the three RL-only pairs decline on TriviaQA, and most RL-only or distillation pairs decline across both tasks [188]. Long-form tests also find more hallucinations for DeepSeek-R1 and QwQ-32B than selected counterparts [20]. These comparisons motivate training hypotheses but do not isolate reasoning: data, rewards, and optimization change together.

A study of post-training separately evaluates access to individual facts and the ability to combine them [192]. Its temporal-chain comparisons are correlational. They support separate reports of factual access and composition, not a universal ranking of training methods.

Trace evidence and interventions. Within one checkpoint, additional computation increases recall, but traces with wrong extracted facts have lower accuracy [45]. Because the facts were observed rather than inserted, difficulty, confidence, or other trace properties may explain that gap.

Interventions expose hidden detector cues, computation spent on decoys [11, 23], unavailable tools, and local edits that fail to repair output [106, 189]. The correction audit is especially informative: accepting a valid edit and recovering the final answer are different outcomes. Its aggregate rates do not establish the probability of recovery after acceptance; S3 preserves their denominators. These tests identify particular mechanisms, not a general trace-length effect.

Audits of visible traces complement these interventions by locating errors: RFC traces repeat unsupported claims [106], graph-coloring traces can invent input edges [56], and RFEval elicits several unfaithful behaviors [55]. Their prevalence estimates depend on the generator, task, and annotation rule.

Evaluator and deployment reports. Grounded-summarization and provider results apply to named systems and dated protocols [12, 119, 128]. Vectara’s successive suites change documents and evaluation; PersonQA follows its system-card protocol. FaithJudge also shows that few-shot

Hypothesis	Proposed test: change → measure			Current evidence
H1 Unchecked claims accumulate	 Generation / verification	-- →	New claims and corrections	Limited / mixed
H2 Rewards favor answering	 Reward arms	-- →	Answer / abstain	Supported synthesis Bounded interventions
H3 Relevance outweighs evidence	 Relevance × support	-- →	Endorse / correct	Limited / mixed
H4 Errors are reused	 Seeded premise	-- →	Later reuse	Direct, one setting Seeded memory attack
H5 Uncertainty changes action	 Signal only	-- →	Verify / abstain / escalate	Supported synthesis Limited transfer

Arrows specify proposed tests, not established causal effects.

Fig. 6. Five hypotheses, their proposed tests, and current evidence. Each row pairs a change with an outcome to measure; dashed arrows do not denote established causal effects. H4 has direct evidence from one seeded memory-attack setting, not an estimate of ordinary error reuse. The tests require the matching controls described in the text. Table 5 specifies the remaining decisive experiments.

Table 5. H1–H5: claims, evidence, and decisive tests. Representative studies support the indicated part of each hypothesis, not every link. The last column gives the remaining experiment.

Hypothesis	Evidence grade	Representative evidence	Decisive test still missing
H1: Unchecked claims accumulate	Limited or mixed	Faulty traces and complexity-related errors [56, 106]	Fix the initial trace; vary unchecked claims, verification, and stopping
H2: Rewards favor answering	Supported synthesis; bounded reward interventions	Reward granularity and calibrated reflection [35, 131]	Randomize reward arms from one base checkpoint; fix data, budget, optimization, and evaluation; use multiple seeds
H3: Relevance outweighs evidence	Limited or mixed	Wrong guidance and sensitivity to supplied evidence [106, 150]	Vary relevance and support for one task using matched wording templates
H4: Errors are reused	Direct, single-setting seeded attack	FARMA: forged reasoning stored and reintroduced [78]	Seed supported, unsupported, and refuted premises; vary independent repair and measure actual reuse
H5: Uncertainty changes action	Supported synthesis; limited transfer	Uncertainty-guided control and the monitoring–control gap [184, 193]	Change only the signal; measure action and useful coverage, then test distribution shift

judge choice changes agreement with human labels [140, 157]. These scores cannot form one continuous ranking.

4.3 Five Mechanism Hypotheses

The evidence grades in Figure 6 follow Section 2.4. Table 5 separates supported links from missing tests; neither single-setting evidence nor synthesis establishes transfer by itself.

H1—unchecked claims accumulate. Longer traces create more opportunities for error, but also for correction. Faulty traces and complexity-related graph errors fit H1 [56, 106]; improved recall and reduced detector visibility complicate that interpretation [23, 45, 204]. A test must count corrections as well as new claims while varying verification and stopping. S4 gives a simple independent-error illustration, not an assumption about real traces.

H2—rewards favor answering. Sparse outcome rewards may favor plausible answers over calibrated abstention. Theory and training studies implicate rare facts, noisy gradients, and incorrect reward optima [77, 88, 175]. Process and outcome rewards yield different errors in small-model mathematics, under an automated judge [131]. Reasoning-finetuned models may also refuse less on unanswerable items or call unavailable tools [81, 153, 189]. Randomized reward arms (Table 5) should measure hallucination, useful coverage, and unnecessary abstention on the same items [35, 178].

H3—relevance outweighs evidence. Following wrong guidance, agreeing with users, and underusing context are compatible with H3 [106, 147, 150], but studies operationalize relevance and support differently. Matched wording templates could vary both factors for one task. Preferential endorsement of relevant unsupported claims would support H3; correction in response to contradictory evidence would instead indicate grounding.

H4—errors are reused. Wrong guidance can recur, and traces containing wrong extracted facts have poorer outcomes [45, 106]. FARMA directly demonstrates forged reasoning stored in memory and reused [78], establishing an attack path rather than its ordinary frequency. Self-feedback, retained feedback, and tool-assisted checks can also correct errors [46, 111, 151]. Testing general reuse requires both seeded premise status and independent repair to vary.

H5—uncertainty changes action. Confidence, hidden states, sparse features, and information flow supply different signals [9, 15, 16, 69, 116]. Controllers combine uncertainty with progress, critique, or trained reflection [66, 165, 184], yet a model may acknowledge contradiction without resolving it [193]. Activation interventions offer ways to isolate the signal [37, 90, 162]. The decisive outcome is changed verification or abstention at useful coverage, including under distribution shift.

4.4 Boundary Conditions for Synthesis

Four variables explain many apparent disagreements. Output detail changes how many claims can be checked, so long- and short-answer rates are not directly comparable [20, 124]. Answer willingness changes the answered subset and can produce non-monotonic budget effects [198]. Evaluator access limits available checks, while checkpoint and training changes may dominate the named reasoning mode [188].

Within-checkpoint recall and between-checkpoint long-form comparisons test different links [20, 45]. Tool interventions identify action failures [189]; observed trace groups suggest intervention targets. Both are informative within scope. Results are excluded when the denominator, source identity, or supported claim cannot be resolved.

Section 5 uses these distinctions to assess when a detection signal can guide verification, repair, or a different release decision.

RQ2 synthesis. Pipeline differences and harmful traces are documented, but no comparison in the 34-source S3 review or 24 reviewed external sources establishes a matched joint adverse effect. The open question concerns a specified intervention, with reasoning and factuality measured jointly under the controls in Section 2.5.

5 From Detection to Intervention

After detecting a suspect claim, the system must decide whether to verify, revise, or withhold the answer. UIPCA identifies the diagnostic target; study design and evidence grades indicate what the supporting studies establish. The method families below connect diagnoses to checks and actions. Interface and model access, timing, and cost determine which interventions are feasible.

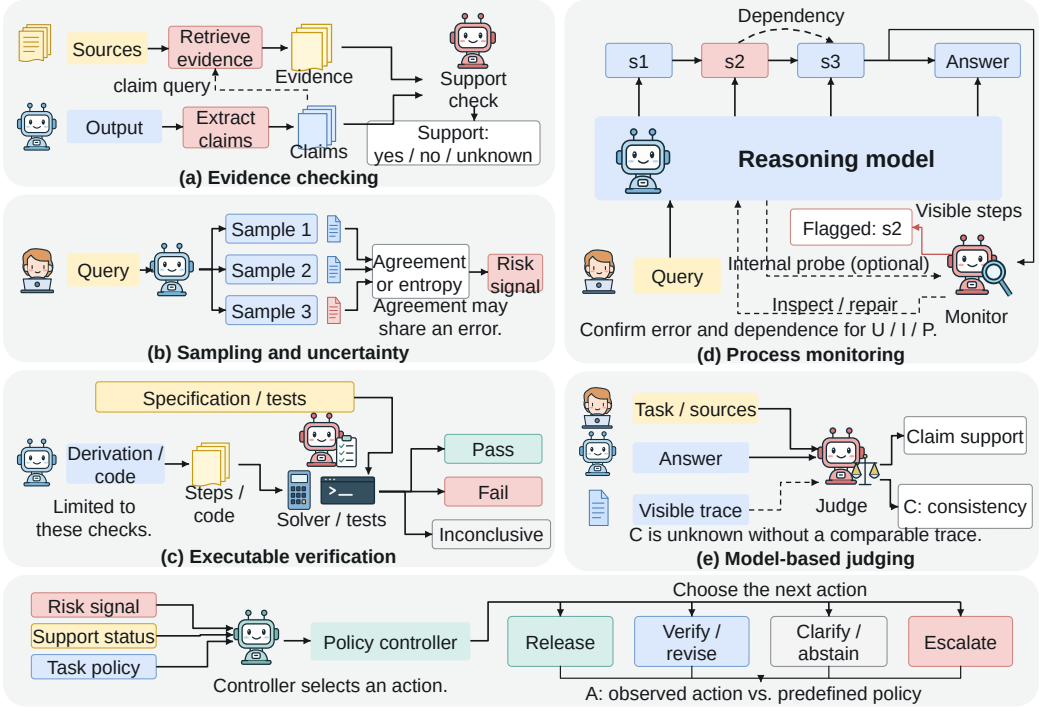


Fig. 7. Complementary detection mechanisms and action control. Panels (a)–(e) illustrate evidence checks, sampling, execution, monitoring, and model-based judging. The controller chooses an action; A audits policy compliance. Table 6 compares the methods.

5.1 Signals, Units, and Actions

Output uncertainty, sample agreement, internal monitors, and external evidence serve different purposes. Semantic Entropy and SelfCheckGPT require repeated generations, but agreement is weak evidence when samples share the same false premise [43, 114]. Hidden-state and sparse-feature probes can flag cases earlier, although their signals often depend on the checkpoint [15, 59, 70, 83]. Claim extraction followed by source checking can provide direct evidence of support, at the cost of extraction, retrieval, and checking [25, 40, 61, 158, 174].

Methods also differ in what they check. FactScore evaluates individual facts, while MiniCheck tests claims against documents [124, 158]. FacTool and SAFE seek external evidence; self-checking instead compares generated outputs [25, 123, 126, 174]. Step detectors target process errors [155], and streaming monitors can trigger stopping or repair during generation [106, 203]. These units affect the available repair: a sentence score may not locate a faulty step, and an internal anomaly cannot supply a missing fact.

Figure 7 shows checks that can be combined within one system. Icons denote functional roles; sample colors in panel (b) distinguish meanings rather than confirmed errors. Table 6 groups methods by the decisions they enable. The complete method catalog, source metrics, and locators are in S5.

Table 6. Detection-to-action comparison. Component mappings identify possible uses of a signal, not a claim that every listed method directly annotates UIPCA.

Signal / Examples	Unit / Covered Component	Interface / Access	Feasible Action	Boundary and Cost
External evidence: CLATTER, RefChecker [40, 61]	Claim; U and source-backed I	All interfaces; source required	Accept, retrieve, revise, abstain	Strong support; extraction and checking cost grows with the number of claims
Uncertainty or learned scoring: Semantic Entropy Probes, FaithLens, UPAIR [83, 152, 184]	Answer/trace; routes checks rather than establishing support or C	White-box probe or learned detector	Verify, resample, reject	Checkpoint and training-domain dependence
Verification and calibrated action: VeriFY, CURE, Latent Critic [3, 103, 165]	Claim/answer; support checks and action selection	Black-box model, labels, or internal critic	Answer, verify, abstain	Coverage depends on boundary estimation and detector transfer
White-box trajectory: TRACED, HARP, Reasoning Denoiser [42, 59, 70]	Step/trajectory; I and P routing	Internal states or trained features	Stop, inspect, escalate	Early but checkpoint-specific; no external factual evidence
Online search/repair: HaluSearch, Token-Guard [24, 203]	Prefix/segment; I and P	Generation-time score and access	Redirect, prune, regenerate	Detection and corrective policy effects are coupled
Executable and grounded checks [72, 97, 110, 132]	Constraint, region, program state; U/I	Evidence, tools, or sandbox	Bind, execute, repair, abstain	The check covers only declared evidence and tested properties
Human or expert review	Consequential claim/decision; all observable components	Structured case packet	Clarify, revise, escalate	Often slow and costly; quality depends on evidence presentation

A fast monitor is most useful when it identifies cases that benefit from a stronger check. Its ranking accuracy alone does not show whether those checks prevent wrong answers or create excessive review work. Detector comparisons and complete-system comparisons therefore answer different questions; S5 gives the conditions needed for each.

Error location guides repair. Case A needs an arithmetic check before reuse; Case B needs a constraint check at Step 8, although an answer–trace comparison can still detect C. PRISM likewise separates memory, instruction, and reasoning failures for diagnosis [179]. For control, the value of localization lies in improving retrieval, repair, or escalation.

A trace can guide computation without faithfully explaining it: chemistry drafts remain useful even when their verbal claims lack support [89]. Conversely, reasoning text may persuade a monitor or lose warning cues after length-penalized training [98, 194]. A routing signal is therefore not itself evidence for a claim.

Access further limits these uses. A provider may run an internal monitor that a user cannot inspect or reproduce. A user-facing summary supports checks on disclosed claims only; omitted process evidence remains $\perp$, meaning insufficient information.

5.2 Intervention Families

Inference-time methods add computation, evidence, or earlier verification. Budget and length controls alter latency [7, 21]. Extra deliberation can improve recall, but its effects on factuality and willingness to answer vary across models [45, 75, 198]. Multi-agent debate adds paths that may share source errors [39]. Answer rate and selective accuracy distinguish reliability gains from increased refusal. Completion, path diversity, and cost help assess extra computation; trace length alone establishes neither repair nor waste.

Independent evidence gives a disputed claim a stronger check than reconsideration alone. Verification, retrieval, and tool-assisted critique can offer such checks [46, 87, 92, 97, 187], although weak retrieval or unavailable tools can introduce errors [189]. Comparing prevention, interruption, rejection, and recovery therefore requires matched evidence and cost budgets.

Early verification can reject a faulty step before reuse. Decoding-time mode switches or verification can interrupt generation [52], while HaluSearch couples detection with search [24]. Clinical critique-and-regenerate methods instead operate on completed answers [93]. They can improve what is released while leaving the original trace unrepaired. Both approaches require an independent final check.

Process supervision moves checking into training. Instead of rewarding only the final answer, it provides feedback on intermediate steps. Labels backed by source evidence can identify unsupported premises (*U*); rule or execution checks can identify invalid transitions (*I*) [28, 94]. Detecting propagation (*P*) requires more: a later step must be shown to depend on a confirmed error. Knowing that both steps are wrong does not establish that connection.

Process/outcome reward comparisons in mathematics and step verification for scientific tools illustrate these differences [131, 199]. Factuality and faithfulness objectives target other units [88, 127, 143]. A step reward can miss a false premise; a factuality verifier can miss an invalid transition. Table 7 relates supervision to these faults.

Table 7. Process supervision and verifier placement by lifecycle target. No verifier type covers all components.

Supervision or verifier	Target	Useful intervention	Characteristic failure
Evidence-backed premise check	<i>U</i>	Reject or retrieve before a premise enters	Open-world evidence gaps; judge shares generator bias
Rule, math, or execution check	<i>I</i>	Reject an invalid local transition	Incomplete checker; valid alternative paths scored as wrong
Dependency-aware reward	<i>P</i>	Stop reuse or invalidate a branch	Missing edges; rewards favor style or late correction rather than preventing reuse
Answer-trace or action verifier	<i>C/A</i>	Compare the answer with the trace; revise, verify, abstain, or escalate	Late detection; exploitable scores, unnecessary refusal, or an unrepaired trace

Final-outcome rewards can improve answers while leaving *U/I/P* unmeasured. Establishing a change in answer-trace consistency (*C*) also requires a trace check. Confidence and abstention objectives instead train uncertainty expression or action choice [3, 103, 178, 195]. Long-form factuality RL balances precision, detail, and relevance [20].

The feedback itself can also be unreliable. Claim-level falsification concentrates checks on decision-relevant content, so its value depends on extraction and verifier quality [183]. Automated labels may inherit judge bias or reject valid alternatives; optimization may exploit the score or favor preferred wording. Evaluating these gains therefore requires independent evidence and an audited subset.

An overall score can also hide differences between outcomes. Reasoning training may amplify deliberative language without correctness-associated behaviors [36]. Changing the reward alters error patterns [131], while penalizing length reduces monitoring cues [98]. Autoformalizers illustrate

a related distinction: a proof may concern a silently repaired statement rather than the original one [30]. Separate checks on reasoning, monitoring, and fidelity to the task reveal what the aggregate gain leaves unresolved.

Attribution to a particular verifier poses a separate question. When data, supervision, and optimization change together, any overall gain concerns the full pipeline (Section 2.5). For deployment, the decisive question is whether the improved check leads to better subsequent answers or actions without excessive refusal or cost.

Other interventions change generation more directly. Decoding methods adjust scores for candidate tokens, while activation methods steer internal representations using signals associated with truthfulness [27, 90]. Knowledge editors such as ROME and MEMIT modify factual associations in model parameters [121, 122]. Multi-hop and sequential tests reveal that an edited fact may not carry through to dependent answers, and unrelated knowledge may also change [53, 200]. The edited answer is therefore only one part of the outcome to assess.

Across domains, the same diagnostic questions call for different evidence and checkers. SPD-Faith separates missed perception from unused evidence; HKVLM checks region binding; Reflect-R1 retrieves video evidence [19, 109, 110]. Joint changes to representation and generation require separate grounding, reasoning, and answer scores [65, 86, 107]. CodeHalu, EvalPlus, and SWE-bench test execution or repair at different units [71, 100, 160]. Passing selected tests establishes neither security nor satisfaction of all requirements.

5.3 Worked Interface-to-Action Comparison

Consider three views of Case A: the full trace, an illustrative summary, and answer only. All show remainder 27 for $1024 \times 1047 \times 1050 \times 1053$ divided by 33, instead of 3. They are views of one completed run, not additional model runs or observed provider summaries.

The full trace exposes $1024 \bmod 33 = 29$ at Step 3; recalculation gives 1, confirming $I_3 = 1$. Later reuse establishes P . Repair must recheck the other residues and dependent products. Agreement with the wrong trace gives $C = 0$, not arithmetic correctness.

A summary that mentions modular arithmetic and remainder 27 omits the reductions needed to identify I/P and τ^* . With answer only, C is also unavailable. Yet independent calculation from the question gives 3 under either restricted view: the result remains checkable even when its origin is hidden.

Under a hypothetical policy requiring independent numerical verification, all views would block 27 and trigger correction; incomplete verification would require abstention. The source case still has $A = \perp$, since it declares no such policy.

Thus localization and answer verification offer distinct benefits. A full trace can reveal where to restart the calculation; external checking can still prevent the wrong answer from being released when the trace is hidden. Whether the additional localization is worth its cost is an empirical question, not an automatic consequence of greater visibility.

5.4 A Calibrated Control Cascade

Hallucination control begins by splitting the answer and visible reasoning into claims, then checking each claim against declared evidence. Claims are supported, unsupported, or unknown. When visible information permits, unsupported claims are traced to their origin and receive component-specific repair. Unknown claims go to retrieval, clarification, abstention, or escalation; they never count as supported. Harm, reversibility, coverage, and cost guide action selection. The subsequent A audit checks policy compliance without changing evidence status.

The practical advantage of a cascade is that expensive checks need not run on every claim. A fast screen can select cases for retrieval, execution, or expert review, while the task determines

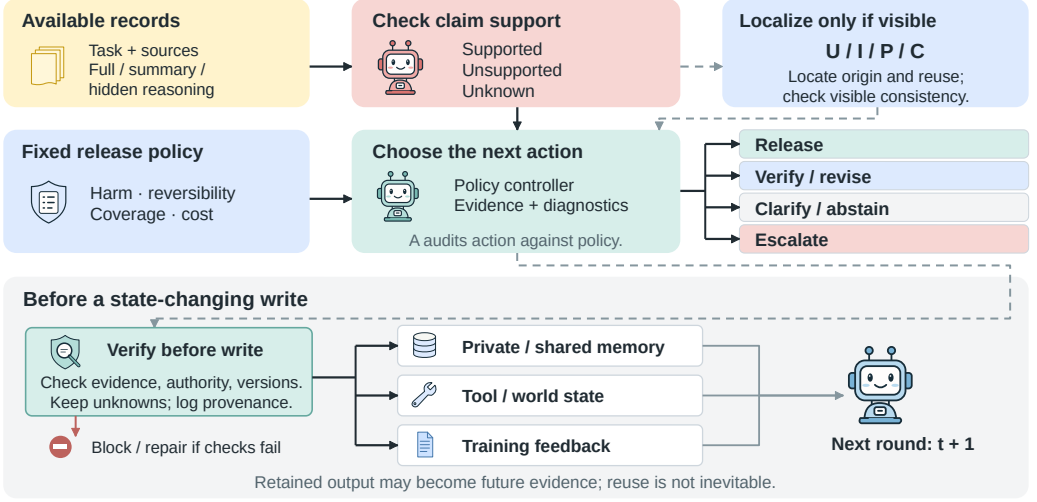


Fig. 8. From claim checks to release and persistent-state control. Claim support and visible-process diagnostics inform a controller governed by a fixed policy. Harm, reversibility, coverage, and cost affect action selection, not evidence status. A audits the chosen action against that policy. State-changing operations also require evidence, write-authority, and version checks. Private or shared memory, tool or world state, and training feedback are parallel destinations. Dashed paths are conditional; retained output may become evidence in a later round, but reuse is not inevitable.

which check can resolve the doubt. A factual dispute may need a source; an arithmetic dispute may need recalculation rather than more generated explanation.

The stages nevertheless cannot be judged in isolation. A screen changes which cases reach the verifier, and a conservative controller can appear reliable simply by escalating almost everything. Two quantities make this trade-off visible: *coverage* is the fraction of requests answered, and *selective risk* is the error rate among those answers. The complete system must be assessed alongside its repair success, workload, and cost. S5 gives the reporting protocol.

RQ3 synthesis. No detector family covers every evidence object, interface, and action. Fast signals can route cases to stronger evidence checks, but their value depends on transfer, false positives, and downstream workload. Consequential release decisions require evaluation of the complete routing and verification process.

6 Evaluation, Deployment, and Research Agenda

6.1 Benchmark Coverage and Joint Metrics

Benchmark families answer different questions. QA and grounding suites assess outcomes [74, 96, 124, 158, 172]; process suites expose or perturb steps [125, 148, 166]; selective suites test confidence or answerability [75, 80, 81]. Domain suites add perception or execution [47, 48, 99, 160]. Table 8 compares their direct evidence and missing companion measurements.

The choice of benchmark determines which diagnosis is possible. An answer score cannot locate the first reasoning error without a trace. Conversely, a step-labeled dataset may reveal an invalid inference without showing whether the system should have answered or abstained. These are complementary measurements, not interchangeable reliability scores.

Domain scores remain conditional. Objects, grounding, and answers do not reveal whether a visual region was used [48, 91, 167]. State prediction, tests, and repair do not establish full requirements or security [47, 71, 100, 160]. Long-context and dialogue tasks additionally need retrieval and binding records.

Memory evaluation extends this comparison beyond the current context. LoCoMo tests long-term conversational recall [112], while LongMemEval includes knowledge updates and abstention [177]. MemoryAgentBench adds incremental learning and selective forgetting [63]. LongMemEval-V2 tests evidence gathered from past web-agent trajectories, including false-premise questions [176]. These tasks assess what memory makes available to a later answer.

HaluMem instead scores extraction, updates, and memory-based QA separately, providing closer access to the operation where content becomes unsupported [17]. MemoryData evaluates component choices, retrieval fidelity, update behavior, and cost [202]. Neither a memory-operation label nor a retrieved source ID by itself establishes dependent reuse under UIPCA.

Table 8. Benchmark families and reporting coverage needed for a release decision. Named resources illustrate each family; protocols and metrics are not interchangeable. Memory governance is separated from answer quality and semantic memory errors.

Family	Primary Unit	Directly Reveals	Required Companion Report
Factual QA / grounding: SimpleQA, TriviaQA, FACTS [22, 74, 172]	Answer or claim–source pair	Outcome factuality or source support	Evidence/evaluator version; extraction rule; answer rate and coverage
Trace / perturbation: FINE-CoT, C2-Faith, RACE [125, 148, 166]	Step, dependency, trace, answer	Selected process errors; coverage depends on suite labels	Segmentation, codebook, boundary cases, interface, unobserved steps
Calibration / answerability: RMCB, VeriFY, AbstentionBench [3, 80, 81]	Confidence, answer, action	Ranking, calibration, answer/abstain behavior	Brier/ECE, selective risk, coverage, unnecessary refusal
Long context / changing evidence: RULER, Lost in the Middle, StreamingQA, HalluDial [58, 101, 105, 108]	Retrieved answer or turn	Outcome under available, positioned, or changing evidence	Retrieval/binding record, evidence time, first-error annotation
Multimodal: POPE, HallusionBench, AMBER [48, 91, 167]	Object, region, claim, answer	Perception/grounding and selected outcome	Coverage, binding, reasoning, answer, abstention as separate fields
Code / software: CodeHalu, CRUXEval, EvalPlus, SWE-bench [47, 71, 100, 160]	Program state, function, repair, patch	Selected execution, repair, or issue-resolution behavior	Environment, tests/coverage, analyzer, security, repository snapshot
Long-term memory: LoCoMo, LongMemEval, MemoryAgentBench, LongMemEval-V2 [63, 112, 176, 177]	History, retrieved context, answer	Recall, updates, forgetting, or environment knowledge; some suites include abstention	History version, evidence-bearing turns, retrieval budget, supported answers and unnecessary refusal
Memory operations: HaluMem, MemoryData [17, 202]	Extracted item, memory update, retrieval result	Semantic memory errors or component-level quality and cost	Before/after state, source support, evaluator, operation failures, downstream outcomes
Shared-memory governance: ArgusFleet [115]	Authorized write/read, supersession, provenance chain	Scope enforcement, state visibility, and conflict handling in the tested service	Identity/scope, versions, concurrency, read path, dependent use, revocation and recovery

6.2 Critical Benchmark Gaps

Three weaknesses recur across the benchmark catalog. **First, outcome-only scoring hides the path.** One score can collapse origin, reuse, answer–trace consistency, and action. It cannot show whether a correct answer relied on an unsupported trace or where a wrong answer began.

Second, many labels depend on an LLM judge. Judge setup can change agreement with audited labels, and surface cues may replace the intended factuality test [140, 157]. FaithformBench tests whether controlled valid and invalid inputs preserve their status [30]. HALLMARK shows that false positives can dominate verifier deployment even when headline recall is high [142].

Errors may be correlated when generator and judge share a model family or training style. Their agreement can then reflect a shared mistake rather than independent confirmation. A judge is useful for routing review, but its reliability has to be checked against evidence outside that agreement.

Memory benchmarks make this separation especially important. HaluMem uses GPT-4o for both memory-based answer generation and scoring, although dataset construction includes human quality checks [17]. Agreement within that pipeline is not an independent factuality guarantee. Judge updates create another ambiguity: a changed score may reflect a changed evaluator rather than better memory. Stable reference cases and independent evidence are needed to distinguish the two.

Third, action coverage is thin. Methods increasingly use signals to answer, stop, revise, or abstain. A method demonstration is not a benchmark for the full policy. Existing tasks rarely cover clarification, retrieval, tool verification, expert escalation, memory-write denial, or rollback. Shared benchmarks provide the least coverage of action-policy evaluation (A). Better benchmarks need process labels, externally checked judges, and complete action and cost records.

6.3 A Lifecycle-Aware Construction Agenda

The missing benchmark is not simply a larger collection of wrong answers. It is a collection that connects an error, the evidence available to detect it, and the action taken in response. Different views of the same run can reveal how much diagnosis is lost when steps are summarized or hidden. Controlled changes to premises or inferences can then test whether a proposed safeguard prevents harm, rather than merely recognizing difficult cases.

Evaluation must also distinguish better answers from fewer answers [35, 80, 103]. Confidence metrics address only part of that distinction. AUROC measures how well a score ranks cases, not whether its numerical confidence is accurate. Brier score and expected calibration error (ECE) assess probability predictions in different ways; S4 gives their definitions and limitations. Similarly, evidence that a source supports a step is not evidence that the step influenced the answer [40, 70, 166].

VeriFY illustrates why metric definitions matter. Its model-relative precision is correctness among released answers; recall compares correct answers with correct answers plus unnecessary abstentions [3]. Recall thus depends on how the protocol establishes the base model’s knowledge, not objective answerability alone. S4 gives the count definitions.

For an actionable benchmark, these measurements need to accompany the same cases. A useful error location should help repair an answer or prevent an unsafe action, not merely improve a step-ranking score. S5 supplies the detailed construction and reporting protocol, including escalation, write denial, and rollback.

6.4 Interface-Aware Release Policy

A release policy decides whether to answer, gather evidence, revise, ask for clarification, abstain, or escalate. It must balance avoiding errors with providing useful answers. Refusing everything avoids releasing a wrong answer but provides no useful coverage; answering everything may exceed the application’s tolerance for harm. The policy’s operating point depends on that tolerance and the capacity for further review.

For illustration, releasing 6,000 of 10,000 requests with 30 wrong answers gives an observed selective risk of 0.5% and coverage of 60%. These estimates meet a maximum risk of 0.6% and

minimum coverage of 55%, but not a maximum risk of 0.2%. Sampling uncertainty and changes in operational data still matter. The thresholds should reflect tolerated harm and useful service, not be chosen after evaluation to make a policy appear successful.

Interface visibility limits which evidence can inform that choice. Full traces allow process checks; summaries allow checks only on disclosed claims. An unavailable process check cannot be assigned zero error. If the risk and coverage requirements cannot both be met with the available evidence and resources, the policy is infeasible. S4 gives the optimization template; S5 gives threshold-selection and uncertainty procedures.

The same signal can justify different actions. A reversible writing error may warrant a warning; a dose recommendation missing patient information may require clarification or clinical review [81, 93]. Passing software tests may still leave security unchecked [71, 100]. Score the action taken: expressed uncertainty is not protection unless it changes unsafe release.

Operational comparisons can otherwise reward extra refusal or unreported computation rather than better control [21, 61, 174, 203]. The relevant outcome is what the complete system releases or does, including the cost of obtaining its evidence.

6.5 From Released Answers to Persistent State

An answer can outlive its turn through memory, shared state, tool effects, or training feedback. Text revision, action rollback, and parameter correction are different recovery operations. Each raises the same question: what evidence is required before a claim persists beyond the current turn?

The return path in Figure 8 requires reuse before repair, not merely retention. Self-Refine supplies self-feedback, Reflexion retains feedback across trials, and CRITIC adds tool-assisted checks [46, 111, 151]; iteration alone establishes no repair. FARMA demonstrates the opposite path: forged reasoning histories are stored and reintroduced [78]. This is the single-setting attack evidence for H4.

Existing memory controls act at different points (Table 9). A-MAC decides which candidate memories to admit, update, or reject using a weighted admission score [196]. Its factual-confidence feature is ROUGE-L overlap with supporting conversation spans, so lexical similarity can influence admission without establishing entailment. TrustMem instead evaluates the transition from old to new memory: a frozen LLM checks coverage, preservation, and support, and its scores guide memory-policy training [186]. This supplies update-level feedback, but support inherited from old memory still needs scrutiny if that memory is already wrong. Admission and consolidation therefore require different checks: whether an item should enter, and whether an edit preserves valid information without adding unsupported claims.

Shared memory adds access and consistency requirements. Governed Memory combines typed properties, entity-scoped retrieval, and organizational policy routing; its reported conflict tests concern sequential writes, not concurrent writers [156]. MemClaw, evaluated through ArgusFleet, attaches scope, supersession, and provenance to shared records [115]. Its evaluation reports a concrete interaction between safeguards: a synchronous near-duplicate gate can reject a contradictory write before the asynchronous conflict detector receives it. A gate intended to prevent redundancy can thus obstruct a legitimate correction. This is evidence about control ordering in one service, not a general rate of memory hallucination.

The distinction changes what an *A* audit must inspect. With private memory, the record can focus on one agent's write decision and its required checks. With shared memory, it must also identify the writer's authority, the versions visible to readers, and the scope of correction or revocation. The definition of *A* remains a violation of a fixed policy; the policy now governs several actors. Provenance can show who wrote or derived a record, but it does not establish that the record is

Table 9. Persistent-memory controls and the evidence they provide. Systems are compared by control point, not by scores from different tasks. Evaluation resources in Table 8 serve a separate role.

System	Control point	Check or evidence	Remaining boundary
A-MAC [196]	Candidate admission and conflict-aware replacement	Weighted utility, overlap-based confidence, novelty, recency, and type prior	Admission precision/recall does not establish semantic truth; overlap is a support proxy
TrustMem [186]	Learned write, revise, and prune behavior	Transition verifier scores coverage, preservation, and faithfulness for reward and preference training	Frozen-LLM judgments depend on the supplied chunk and prior memory; no independent truth guarantee
Governed Memory [156]	Typed extraction, entity-scoped retrieval, policy routing	Heuristic quality gates, schema checks, provenance metadata, and bounded retrieval	Self-evaluation; sequential conflict tests leave concurrent multi-agent writes unresolved
MemClaw / ArgusFleet [115]	Scoped access, supersession, and controlled propagation	Service traces test authorized visibility, conflict handling, and provenance reconstruction	One-service self-evaluation; admission can pre-empt correction; visibility is not dependent reasoning use

true. Nor does visibility to another agent establish P : the later claim must depend on a confirmed error. A valid, unauthorized disclosure is a governance failure even when no hallucination occurs.

These distinctions also clarify recovery. Denying a write prevents one form of persistence; superseding a record changes which version should be retrieved. Neither operation alone repairs summaries, tool actions, or feedback already derived from that record. A recovery evaluation should therefore follow affected consumers after correction, not stop when the storage API acknowledges an update. The memory benchmarks in Table 8 supply parts of this evidence, while a joint test of semantic correction, authorized access, and downstream recovery remains necessary.

The useful quantity here is not how often a claim is stored, but how many later consumers reuse it before independent repair. We denote that expected count by $\mathcal{R}_{H,t}$. A consumer may be another agent, a later reasoning round, or another operation that uses the retained claim. Its identity and the observation horizon must be fixed for the count to be meaningful.

For a simple illustration, suppose three consumers each have a 50% chance of selecting a claim, and a selected claim has a 20% chance of repair before use. The expected number of unrepaired uses is then $3 \times 0.5 \times 0.8 = 1.2$. Raising the conditional repair probability to 60% reduces it to 0.6. These are hypothetical values, not observed rates. S4 gives the general definition, which also accommodates different reuse and repair probabilities across consumers. S5 describes how to compare checks before writing with correction after reuse.

6.6 Experiments That Close the Open Links

Three open questions follow from this synthesis. The first concerns the apparent trade-off between reasoning and reliability. Extra computation may help when it includes verification, while abstention rewards may improve reported accuracy chiefly by reducing coverage. The matched designs in Section 2.5 can separate these explanations. Their purpose is to identify which intervention changes which outcome, not to rank “more reasoning” as universally better or worse.

The second concerns the practical value of localization. Does knowing the error’s location improve repair beyond a response-level warning? Comparing these signals on the same cases would test UIPCA’s proposed diagnostic benefit more directly than agreement with a label alone. Restricted interfaces are important here: a useful system must act on the evidence it can access, without filling hidden fields with inferred certainty.

The third concerns transfer and recovery. A safeguard can succeed on one task because it has access to decisive evidence that is unavailable elsewhere. Translation may change entities, negation,

units, or evidential strength [49, 159]; context position and visual grounding can change which evidence is used [10, 99, 101, 109]. In code, a passing result remains conditional on requirements and tests [41, 47, 71, 100]. Transfer therefore concerns the adequacy of the check, not just the model’s score in a new domain.

Recovery also depends on timing. Correcting a premise before reuse and repairing a final answer after reuse solve different problems. Local edit acceptance may leave the outcome unchanged [106], and retrieval or tools may move the first error rather than remove it [92, 189]. Comparing early checks, branch interruption, and later correction under common conditions would show where a safeguard still prevents harm. Domain shifts, adversarial evidence, and unavailable tools test whether that benefit survives beyond the original setting [3, 88, 103, 178]. S3 and S5 preserve the experimental controls and statistical procedures for these questions.

RQ4 synthesis. Answer quality, visible process evidence, coverage, action, and cost require separate measurements. Shared benchmarks still cover few clarification, escalation, persistent-write, and rollback decisions. Policy evaluation must therefore specify a risk–coverage operating point and extend beyond the displayed answer to consequential writes.

7 Conclusion

UIPCA separates unsupported premises and invalid inferences, dependent reuse, visible answer–trace consistency, and action-policy failure. These are different diagnostic questions, not sequential stages. Text-centered evidence and selected transfer cases show why repair should follow the earliest useful observable check. Hidden process evidence remains unknown; the record does not infer deceptive intent.

Pipeline differences and harmful traces do not establish that reasoning generally lowers factual reliability. No source in the 34-entry S3 review or the 24-source external follow-up meets the specified matched joint adverse-effect criterion. This bounded finding leaves the causal question open. H1–H5 identify tests that separate the unresolved links while measuring action, coverage, and cost.

Answer factuality, process support, confidence, and action need separate measurements. Their joint evaluation distinguishes better evidence from increased refusal or computation. Consequential outputs also require checks before they enter memory, shared state, tools, or training feedback. Verification and provenance can support writes; unresolved content may need quarantine, and harmful writes may need rollback. Reliability concerns both the current answer and the state inherited by later rounds.

References

- [1] Aisha Alansari and Hamzah Luqman. 2025. Large Language Models Hallucination: A Comprehensive Survey. [cs.CL] <https://arxiv.org/abs/2510.06265>
- [2] Andres Algaba, Francesca Carlon, Lynn Delcon, Marthe Ballon, Bert Verbruggen, and Vincent Ginis. 2026. How Much Does a Reasoning Summary Reveal? An Observability Ladder for Large Language Models. [cs.LG] <https://arxiv.org/abs/2608.02089>
- [3] Enes Altinisik, Masoomali Fatehikia, Fatih Deniz, Nadir Durrani, Majd Hawasly, Mohammad Raza, and Husrev Taha Sencar. 2026. Do I Really Know? Learning Factual Self-Verification for Hallucination Reduction. [cs.AI] <https://arxiv.org/abs/2602.02018>
- [4] Avinash Anand, Mahisha Ramesh, Avni Mittal, Ashutosh Kumar, Rishitej Reddy Vyalla, Erik Cambria, Zhengkui Wang, Timothy Liu, Aik Beng Ng, Simon See, and Rajiv Ratn Shah. 2026. The Periodic Table of LLM Reasoning: A Structured Survey of Reasoning Paradigms, Methods, and Failure Modes. [cs.CL] <https://arxiv.org/abs/2606.11470>
- [5] Suramya R. Angdembay, Dikshant Aryal, and Nick Rahimi. 2026. Two Regimes of Chain-of-Thought Unfaithfulness: Metric-Based Detection Fails Where Models Are Wrong. [cs.CL] <https://arxiv.org/abs/2607.23458>
- [6] Anonymous ACL submission. 2026. Mirages of Logic: A Survey of Chain-of-Thought Reasoning Hallucinations. Under review; public version modified 7 June 2026. <https://openreview.net/forum?id=sINjRFaFkE>

- [7] Anthropic. 2025. Claude’s extended thinking. <https://www.anthropic.com/news/visible-extended-thinking>
- [8] Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooan Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-Thought Reasoning In The Wild Is Not Always Faithful. [cs.AI] <https://arxiv.org/abs/2503.08679>
- [9] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It’s Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 967–976. [cs.CL] doi:10.18653/v1/2023.findings-emnlp.68
- [10] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. [cs.CL] <https://arxiv.org/abs/2308.14508>
- [11] Oluwanifemi Bamgbose, Masoud Hashemi, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, and Vikas Yadav. 2026. DNR Bench: Benchmarking Over-Reasoning in Reasoning LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence* 40, 44 (2026), 37240–37248. [cs.LG] doi:10.1609/aaai.v40i44.41055
- [12] Forrest Bao. 2025. DeepSeek-R1 hallucinates more than DeepSeek-V3. Legacy HHEM 2.1 grounded-summarization evaluation. <https://www.vectara.com/blog/deepseek-r1-hallucinates-more-than-deepseek-v3>
- [13] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating Open-Vocabulary Caption Hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 22680–22698. [cs.CV] doi:10.18653/v1/2024.emnlp-main.1263
- [14] Krischan Braitsch, Laura K. Schmalbrock, Theresa Weltermann, Andrew F. Berdel, Isabella Miller, Kai Tran, Michael Heider, Sabrina Kraus, Florian Bassermann, Jacqueline Lammert, Sebastian Ziegelmayr, Marcus Makowski, Lisa C. Adams, and Keno K. Bressem. 2026. Information-Seeking Failures of Large Language Models in Agentic Clinical Reasoning. [cs.AI] <https://arxiv.org/abs/2607.10275>
- [15] Boshui Chen, Zhaoxin Fan, Ke Wang, Zhiying Leng, Faguo Wu, Hongwei Zheng, Yifan Sun, and Wenjun Wu. 2026. HalluSAE: Detecting Hallucinations in Large Language Models via Sparse Auto-Encoders. [cs.CL] <https://arxiv.org/abs/2604.16430>
- [16] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs’ Internal States Retain the Power of Hallucination Detection. [cs.CL] <https://arxiv.org/abs/2402.03744>
- [17] Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchu Li, Feiyu Xiong, and Zhiyu Li. 2025. HaluMem: Evaluating Hallucinations in Memory Systems of Agents. <https://arxiv.org/abs/2511.03506v3>
- [18] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026. Towards reasoning era: a survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences* 69, 6 (2026), 161101. doi:10.1007/s11432-025-4665-8
- [19] Shuimu Chen, Yuteng Chen, Yuanshen Guan, Zebang Cheng, Zeyu Zhang, Shengqian Qin, Bin Xia, Jiaran Li, Wenming Yang, and Fei Ma. 2026. Reflect-R1: Evidence-Driven Reflection for Self-Correction in Long Video Understanding. [cs.CV] <https://arxiv.org/abs/2606.27922>
- [20] Xilun Chen, Iliia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen tau Yih. 2025. Learning to Reason for Factuality. [cs.CL] <https://arxiv.org/abs/2508.05618>
- [21] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs. [cs.CL] <https://arxiv.org/abs/2412.21187>
- [22] Aileen Cheng, Alon Jacovi, Amir Globerson, Ben Golan, Charles Kwong, Chris Alberti, Connie Tao, Eyal Ben-David, Gaurav Singh Tomar, Lukas Haas, Yonatan Bitton, Adam Bloniarz, Aijun Bai, Andrew Wang, Anfal Siddiqui, Arturo Bajuelos Castillo, Aviel Atias, Chang Liu, Corey Fry, Daniel Balle, Deepanway Ghosal, Doron Kukliansky, Dror Marcus, Elena Gribovskaya, Eran Ofek, Honglei Zhuang, Itay Laish, Jan Ackermann, Lily Wang, Meg Risdal, Megan Barnes, Michael Fink, Mohamed Amin, Moran Ambar, Natan Potikha, Nikita Gupta, Nitzan Katz, Noam Velan, Ofir Roval, Ori Ram, Polina Zablotskaia, Prathamesh Bang, Priyanka Agrawal, Rakesh Ghiya, Sanjay Ganapathy, Simon Baumgartner, Sofia Erell, Sushant Prakash, Thibault Sellam, Vikram Rao, Xuanhui Wang, Yaroslav Akulov, Yulong Yang, Zhen Yang, Zhixin Lai, Zhongru Wu, Anca Dragan, Avinatan Hassidim, Fernando Pereira, Slav Petrov, Srinivasan Venkatachary, Tulsee Doshi, Yossi Matias, Sasha Goldshtein, and Dipanjan Das. 2025. The FACTS Leaderboard: A Comprehensive Benchmark for Large Language Model Factuality. [cs.CL] <https://arxiv.org/abs/2512.10791>
- [23] Jiahao Cheng, Tiancheng Su, Jia Yuan, Guoxiu He, Jiawei Liu, Xinqi Tao, Jingwen Xie, and Huaxia Li. 2025. Chain-of-Thought Prompting Obscures Hallucination Cues in Large Language Models: An Empirical Evaluation. [cs.CL] <https://arxiv.org/abs/2506.17088>
- [24] Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking. [cs.CL] <https://arxiv.org/abs/2501.01306>
- [25] I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task

- and Multi-Domain Scenarios. [cs.CL] <https://arxiv.org/abs/2307.13528>
- [26] James Chua and Owain Evans. 2025. Are DeepSeek R1 And Other Reasoning Models More Faithful? [cs.LG] <https://arxiv.org/abs/2501.08156>
- [27] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. [cs.CL] <https://arxiv.org/abs/2309.03883>
- [28] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. [cs.LG] <https://arxiv.org/abs/2110.14168>
- [29] Chrisanna Cornish and Anna Rogers. 2025. Examining the Faithfulness of Deepseek R1’s Chain-of-Thought Reasoning. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025)*. Association for Computational Linguistics, Mumbai, India, 11–19. doi:10.18653/v1/2025.chomps-main.2
- [30] Rob Cornish, Iacopo Ghinassi, Po-Hung Yeh, Shuqi Liu, Qiyuan Xu, Haoxuan Yin, Dominik Wagner, Wenda Li, Yee Whye Teh, and Luke Ong. 2026. FaithformBench: Benchmarking Faithfulness of Mathematical Chain-of-Thought Autoformalisation. [cs.CL] <https://arxiv.org/abs/2608.10916>
- [31] Manuel Cossio. 2025. A comprehensive taxonomy of hallucinations in Large Language Models. [cs.CL] <https://arxiv.org/abs/2508.01781>
- [32] Antonia Creswell and Murray Shanahan. 2022. Faithful Reasoning Using Large Language Models. [cs.AI] <https://arxiv.org/abs/2208.14271>
- [33] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–93. [cs.CL] doi:10.1093/jla/laae003
- [34] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8493–8502. [cs.CL] doi:10.18653/v1/2022.acl-long.581
- [35] Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. [cs.LG] <https://arxiv.org/abs/2507.16806>
- [36] Jean de Dieu Nyandwi, Leena Mathur, Yonatan Bisk, Robert Hawkins, and Graham Neubig. 2026. Amplified Does Not Mean Predictive: Reasoning Behaviors in Thinking Models. [cs.CL] <https://arxiv.org/abs/2608.13760>
- [37] Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. 2026. Qwen-Scope: Turning Sparse Features into Development Tools for Large Language Models. [cs.CL] <https://arxiv.org/abs/2605.11887>
- [38] Qihua Dong, Yitian Zhang, Huimin Zeng, Yizhou Wang, Jianglin Lu, Kuo Yang, and Yun Fu. 2026. Revealing the Seen, Imagining the Beyond: A Survey of Image-Grounded Chain-of-Thought Reasoning in Multimodal LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, San Diego, California, United States, 45055–45070. doi:10.18653/v1/2026.acl-long.2087
- [39] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. [cs.CL] <https://arxiv.org/abs/2305.14325>
- [40] Ron Eliav, Arie Cattan, Eran Hirsch, Shahaf Bassan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. 2025. CLATTER: Comprehensive Entailment Reasoning for Hallucination Detection. [cs.CL] <https://arxiv.org/abs/2506.05243>
- [41] Mahdi Erfanian, Nelson Daniel Troncoso, Aashna Garg, Amabel Gale, Xiaoyu Liu, Pareesa Ameneh Golnari, and Shengyu Fu. 2026. Delulu: A Verified Multi-Lingual Benchmark for Code Hallucination Detection in Fill-in-the-Middle Tasks. [cs.LG] <https://arxiv.org/abs/2605.07024>
- [42] Junlin Fang, Do Nguyen-Thanh, Xiaogang Xu, Zhen Fang, and Sean Du. 2026. Reasoning Denoiser: Denoising Reasoning Traces for Hallucination Detection in Large Reasoning Models. [cs.AI] <https://arxiv.org/abs/2607.22098>
- [43] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630. doi:10.1038/s41586-024-07421-0
- [44] Kartik Garg, Shourya Mishra, Kartikeya Sinha, Ojaswi Pratap Singh, Ayush Chopra, Kanishk Rai, Ammar Sheikh, Raghav Maheshwari, Aman Chadha, Vinija Jain, and Amitava Das. 2025. Alignment Faking - the Train -> Deploy Asymmetry: Through a Game-Theoretic Lens with Bayesian-Stackelberg Equilibria. [cs.AI] <https://arxiv.org/abs/2511.17937>
- [45] Zorik Gekhman, Roei Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. 2026. Thinking to Recall: How Reasoning Unlocks Parametric Knowledge in LLMs. [cs.CL] <https://arxiv.org/abs/2603.09906>
- [46] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing. [cs.CL] <https://arxiv.org/abs/2305.11738>

- [47] Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. 2024. CRUXEval: A Benchmark for Code Reasoning, Understanding and Execution. [cs.SE] <https://arxiv.org/abs/2401.03065>
- [48] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 14375–14385. [cs.CV] doi:10.1109/cvpr52733.2024.01363
- [49] Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics* 11 (2023), 1500–1517. [cs.CL] doi:10.1162/tacl_a_00615
- [50] Dadi Guo, Jiayu Liu, Zhiyuan Fan, Zhitao He, Haoran Li, Yuxin Li, Yumeng Wang, and Yi R. Fung. 2025. Mathematical Proof as a Litmus Test: Revealing Failure Modes of Advanced Large Reasoning Models. [cs.AI] <https://arxiv.org/abs/2506.17114>
- [51] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shironi Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiasi Li, Jingchang Chen, Jingyang Yuan, Jinzhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Hu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (2025), 633–638. [cs.CL] doi:10.1038/s41586-025-09422-z
- [52] YiQiu Guo, Yuchen Yang, Zhe Chen, Pingjie Wang, Yusheng Liao, Ya Zhang, Yanfeng Wang, and Yu Wang. 2025. DSVD: Dynamic Self-Verify Decoding for Faithful Generation in Large Language Models. [cs.CL] <https://arxiv.org/abs/2503.03149>
- [53] Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model Editing at Scale leads to Gradual and Catastrophic Forgetting. [cs.CL] <https://arxiv.org/abs/2401.07453>
- [54] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2317967121. [cs.CL] doi:10.1073/pnas.2317967121
- [55] Yunseok Han, Jejoon Lee, and Jaeyoung Do. 2026. RFEval: Benchmarking Reasoning Faithfulness under Counterfactual Reasoning Intervention in Large Reasoning Models. [cs.AI] <https://arxiv.org/abs/2602.17053>
- [56] Alex Heyman and Joel Zylberberg. 2025. Reasoning Large Language Model Errors Arise from Hallucinating Critical Problem Features. [cs.LG] <https://arxiv.org/abs/2505.12151>
- [57] Zheng Yi Ho, Siyuan Liang, and Dacheng Tao. 2026. Understanding Hallucinations in Large Visual and Language Models. *Comput. Surveys* 58, 13, Article 332 (2026), 36 pages. [cs.CV] doi:10.1145/3811409
- [58] Cheng-Ping Hsieh, Siman Sun, Samuel Kriman, Shantanu Acharya, Dima Rekeshe, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What’s the Real Context Size of Your Long-Context Language Models? [cs.CL] <https://arxiv.org/abs/2404.06654>
- [59] Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. 2025. HARP: Hallucination Detection via Reasoning Subspace Projection. [cs.CL] <https://arxiv.org/abs/2509.11536>
- [60] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model. [cs.LG] <https://arxiv.org/abs/2503.24290>

- [61] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. [cs.CL] <https://arxiv.org/abs/2405.14486>
- [62] Yi Hu, Jiaqi Gu, Ruxin Wang, Zijun Yao, Hao Peng, Xiaobao Wu, Jianhui Chen, Muhan Zhang, and Liangming Pan. 2026. Towards a Mechanistic Understanding of Large Reasoning Models: A Survey of Training, Inference, and Failures. [cs.CL] <https://arxiv.org/abs/2601.19928>
- [63] Yuanzhe Hu, Yu Wang, and Julian McAuley. 2026. Evaluating Memory in LLM Agents via Incremental Multi-Turn Interactions. In *The Fourteenth International Conference on Learning Representations*. <https://arxiv.org/abs/2507.05257v4>
- [64] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43, 2, Article 42 (2025), 55 pages. doi:10.1145/3703155
- [65] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 13418–13427. [cs.CV] doi:10.1109/cvpr52733.2024.01274
- [66] Zhengze Huang, Luyang Yu, Di Hong, Xinzhe Huang, Wanyu Lin, Zhixuan Chu, Zhan Qin, and Tianhang Zheng. 2026. REIN: Bridging the Gap between Reasoning and Reliability via Reflection and Abstention Alignment. [cs.AI] <https://arxiv.org/abs/2608.07931>
- [67] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermy, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. [cs.CR] <https://arxiv.org/abs/2401.05566>
- [68] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38. [cs.CL] doi:10.1145/3571730
- [69] Jinghan Jia, Joe Benton, and Eric Easley. 2026. Faithfulness as Information Flow: Evaluating and Training Faithful Chain-of-Thought Reasoning. [cs.LG] <https://arxiv.org/abs/2605.24286>
- [70] Xinyan Jiang, Ninghao Liu, Di Wang, and Lijie Hu. 2026. Beyond Scalars: Evaluating and Understanding LLM Reasoning via Geometric Progress and Stability. [cs.AI] <https://arxiv.org/abs/2603.10384>
- [71] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? [cs.CL] <https://arxiv.org/abs/2310.06770>
- [72] Zhensheng Jin, Xin Dai, Zhenghao Liu, Chaojun Xiao, Huiyuan Xie, Yu Gu, Ge Yu, and Maosong Sun. 2026. REFACT: Adaptive Fact Restatement for Compact and Faithful Chain-of-Thought Reasoning. [cs.CL] <https://arxiv.org/abs/2607.20833>
- [73] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 5042–5063. [cs.CV] doi:10.18653/v1/2024.findings-emnlp.290
- [74] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. [cs.CL] doi:10.18653/v1/P17-1147
- [75] Haeji Jung and Hila Gonen. 2026. PhantomBench: Benchmarking the Non-existential Threat of Language Models. [cs.CL] <https://arxiv.org/abs/2606.11105>
- [76] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. [cs.CL] <https://arxiv.org/abs/2207.05221>
- [77] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2026. Evaluating large language models for accuracy incentivizes hallucinations. *Nature* 653, 8116 (2026), 1047–1051. doi:10.1038/s41586-026-10549-w
- [78] Neeraj Karamchandani, Piyush Nagasubramaniam, Sencun Zhu, and Dinghao Wu. 2026. Your Agent’s Memories Are Not Its Own: Forged Reasoning Attacks on LLM Agent Memory and Defenses. [cs.CR] <https://arxiv.org/abs/2607.05029>

- [79] Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent AI model. <https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/>
- [80] Reza Khanmohammadi, Erfan Miah, Simerjot Kaur, Ivan Brugere, Charese H. Smiley, Kundan Thind, and Mohammad M. Ghassemi. 2026. How Reliable are Confidence Estimators for Large Reasoning Models? A Systematic Benchmark on High-Stakes Domains. [cs.CL] <https://arxiv.org/abs/2601.08134>
- [81] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. In *Advances in Neural Information Processing Systems*, Vol. 38. Curran Associates, Inc., Red Hook, NY, 51 pages. [cs.AI] <https://arxiv.org/abs/2506.09038>
- [82] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models Are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), New Orleans, LA, USA, 22199–22213. [cs.CL] doi:10.52202/068431-1613
- [83] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. [cs.CL] <https://arxiv.org/abs/2406.15927>
- [84] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring Faithfulness in Chain-of-Thought Reasoning. [cs.AI] <https://arxiv.org/abs/2307.13702>
- [85] Yunseo Lee, John Youngeun Song, Dongsun Kim, Jindae Kim, Mijung Kim, and Jaechang Nam. 2025. Hallucination by Code Generation LLMs: Taxonomy, Benchmarks, Mitigation, and Challenges. [cs.SE] <https://arxiv.org/abs/2504.20799>
- [86] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 13872–13882. [cs.CV] doi:10.1109/cvpr52733.2024.01316
- [87] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. [cs.CL] <https://arxiv.org/abs/2005.11401>
- [88] Junyi Li and Hwee Tou Ng. 2025. Reasoning Models Hallucinate More: Factuality-Aware Reinforcement Learning for Large Reasoning Models. [cs.CL] <https://arxiv.org/abs/2505.24630>
- [89] Jiatong Li, Yuxuan Ren, Weida Wang, Xiaoyong Wei, and Yatao Bian. 2026. Chemical Chain-of-Thought Functions as a Hallucination-Prone Molecular Scratchpad. [cs.CE] <https://arxiv.org/abs/2607.20935>
- [90] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems 36*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), New Orleans, LA, USA, 41451–41530. [cs.LG] doi:10.52202/075280-1797
- [91] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. [cs.CV] <https://arxiv.org/abs/2305.10355>
- [92] Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, Qinyuan Cheng, Bo Wang, Yining Zheng, Yuxin Wang, Zhangyue Yin, and Xipeng Qiu. 2025. R3-RAG: Learning Step-by-Step Reasoning and Retrieval for LLMs via Reinforcement Learning. [cs.CL] <https://arxiv.org/abs/2505.23794>
- [93] Yiming Liao, Zeno Franco, Jose Eduardo Lizarraga Mazaba, and Keke Chen. 2026. Med-HEAL: Analyzing and Mitigating Hallucinations in Medical LLMs with Hallucination-Aware In-Context Learning. [cs.CL] <https://arxiv.org/abs/2606.01301>
- [94] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s Verify Step by Step. [cs.LG] <https://arxiv.org/abs/2305.20050>
- [95] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. [cs.CL] <https://arxiv.org/abs/2205.14334>
- [96] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3214–3252. [cs.CL] doi:10.18653/v1/2022.acl-long.229
- [97] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive Verification of Chain-of-Thought Reasoning. [cs.CL] <https://arxiv.org/abs/2306.03872>
- [98] Bryce Little. 2026. Length Penalties Make Chain-of-Thought Less Monitorable. [cs.AI] <https://arxiv.org/abs/2607.09786>
- [99] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More Thinking, Less Seeing? Assessing Amplified Hallucination in Multimodal Reasoning Models. [cs.CL] <https://arxiv.org/abs/2505.21523>

- [100] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. [cs.SE] <https://arxiv.org/abs/2305.01210>
- [101] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. [cs.CL] doi:10.1162/tacl_a_00638
- [102] Xinyi Liu, Yuting Lu, and Shunping Wei. 2026. Survey on Hallucination in Reasoning Large Language Model: Evaluation, Taxonomy, Intervention, and Open Issues. *Data Intelligence* 8 (2026), 20250131. doi:10.3724/2096-7004.di.2025.0131
- [103] Xin Liu and Lu Wang. 2026. Think Through Uncertainty: Improving Long-Form Generation Factuality via Reasoning Calibration. [cs.CL] <https://arxiv.org/abs/2604.12046>
- [104] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding R1-Zero-Like Training: A Critical Perspective. [cs.LG] <https://arxiv.org/abs/2503.20783>
- [105] Adam Liška, Tomáš Kočíský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d’Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models. [cs.CL] <https://arxiv.org/abs/2205.11388>
- [106] Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing Meta-Cognitive Hallucinations in Reasoning Large Language Models. [cs.CY] <https://arxiv.org/abs/2505.13143>
- [107] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. [cs.CL] <https://arxiv.org/abs/2209.09513>
- [108] Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. HalluDial: A Large-Scale Benchmark for Automatic Dialogue-Level Hallucination Evaluation. [cs.CL] <https://arxiv.org/abs/2406.07070>
- [109] Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. 2026. SPD-Faith Bench: Diagnosing and Improving Faithfulness in Chain-of-Thought for Multimodal Large Language Models. [cs.CV] <https://arxiv.org/abs/2602.07833>
- [110] Bo Ma. 2026. HKVLM: Faithful Query–Region Binding for Frozen-Detector Visual Grounding. [cs.CV] <https://arxiv.org/abs/2606.28862>
- [111] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. [cs.CL] <https://arxiv.org/abs/2303.17651>
- [112] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 13851–13870. doi:10.18653/v1/2024.acl-long.747
- [113] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 9802–9822. [cs.CL] doi:10.18653/v1/2023.acl-long.546
- [114] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. [cs.CL] <https://arxiv.org/abs/2303.08896>
- [115] Yanki Margalit, Nurit Cohen-Inger, Erni Avram, Ran Taig, and Oded Margalit. 2026. Governed Shared Memory for Multi-Agent LLM Systems. <https://arxiv.org/abs/2606.24535v1>
- [116] Samuel Marks and Max Tegmark. 2023. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. [cs.AI] <https://arxiv.org/abs/2310.06824>
- [117] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1906–1919. [cs.CL] doi:10.18653/v1/2020.acl-main.173
- [118] Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. 2025. Reasoning about Uncertainty: Do Reasoning Models Know When They Don’t Know? [cs.AI] <https://arxiv.org/abs/2506.18183>
- [119] Ofer Mendelevitch. 2025. Introducing the Next Generation of Vectara’s Hallucination Leaderboard. <https://www.vectara.com/blog/introducing-the-next-generation-of-vectaras-hallucination-leaderboard>
- [120] Shreya Mendi and Brinnae Bent. 2026. Why2Speak: Faithful Reasoning for Abstaining Action Policies. [cs.AI] <https://arxiv.org/abs/2608.20670>
- [121] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation,

- Inc. (NeurIPS), New Orleans, LA, USA, 17359–17372. [cs.CL] doi:10.52202/068431-1262
- [122] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-Editing Memory in a Transformer. [cs.CL] <https://arxiv.org/abs/2210.07229>
- [123] Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. SelfCheck: Using LLMs to Zero-Shot Check Their Own Step-by-Step Reasoning. [cs.AI] <https://arxiv.org/abs/2308.00436>
- [124] Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FACTScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. [cs.CL] <https://arxiv.org/abs/2305.14251>
- [125] Avni Mittal and Rauno Arike. 2026. C2-Faith: Benchmarking LLM Judges for Causal and Coverage Faithfulness in Chain-of-Thought Reasoning. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*. Association for Computational Linguistics, San Diego, California, USA, 678–691. [cs.CL] doi:10.18653/v1/2026.gem-main.62
- [126] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. [cs.CL] <https://arxiv.org/abs/2305.15852>
- [127] Shuo Nie, Hexuan Deng, Chao Wang, Ruiyu Fang, Xuebo Liu, Shuangyong Song, Yu Li, Min Zhang, and Xuelong Li. 2026. Stop Rewarding Hallucinated Steps: Faithfulness-Aware Step-Level Reinforcement Learning for Small Reasoning Models. [cs.CL] <https://arxiv.org/abs/2602.05897>
- [128] OpenAI. 2025. OpenAI o3 and o4-mini System Card. PersonQA results in Section 3.3. <https://openai.com/index/o3-o4-mini-system-card/>
- [129] OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Bohan Zhang, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Wang, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukas Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiye Zheng, Wenda Zhou, Wenting Zhan, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. OpenAI o1 System Card. [cs.AI] <https://arxiv.org/abs/2412.16720>
- [130] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow

- Instructions with Human Feedback. In *Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), New Orleans, LA, USA, 27730–27744. [cs.CL] doi:10.52202/068431-2011
- [131] Anagha Radhakrishna Palandye, Rebecca Glick, and Osheen Kaul. 2026. Reward Granularity in RLVR: Comparing Process and Outcome Reward Structures for Mathematical Reasoning in Small Language Models. [cs.LG] <https://arxiv.org/abs/2607.02869>
- [132] Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 3806–3824. doi:10.18653/v1/2023.findings-emnlp.248
- [133] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 15012–15032. doi:10.18653/v1/2024.findings-emnlp.882
- [134] Ahmad Pesaranghader and Erin Li. 2026. Hallucination Detection and Mitigation in Large Language Models. [cs.AI] <https://arxiv.org/abs/2601.09929>
- [135] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. [cs.CL] doi:10.18653/v1/D19-1250
- [136] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki Van Stein, and Thomas Bäck. 2026. Multi-Step Reasoning with Large Language Models, a Survey. *Comput. Surveys* 58, 6 (2026), 1–35. [cs.AI] doi:10.1145/3774896
- [137] Vishal Pramanik, Maisha Maliha, Nathaniel D. Bastian, Alvaro Velasquez, Susmit Jha, and Sumit Kumar Jha. 2026. NSF-CoT: Neuro-Symbolic Formal Verification of Chain-of-Thought Faithfulness in Contextual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, San Diego, California, United States, 10645–10663. doi:10.18653/v1/2026.findings-acl.516
- [138] Zhe Qian, Yanbiao Ma, Zhuohan Ouyang, Zhonghua Wang, Zhongxing Xu, Fei Luo, Xinyu Liu, Zongyuan Ge, Yike Guo, and Jungong Han. 2026. Cognitive Pivot Points and Visual Anchoring: Unveiling and Rectifying Hallucinations in Multimodal Reasoning Models. [cs.AI] <https://arxiv.org/abs/2604.10219>
- [139] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Question Decomposition Improves the Faithfulness of Model-Generated Reasoning. [cs.CL] <https://arxiv.org/abs/2307.11768>
- [140] Sanjana Ramprasad and Byron C. Wallace. 2024. Do Automatic Factuality Metrics Measure Factuality? A Critical Evaluation. [cs.CL] <https://arxiv.org/abs/2411.16638>
- [141] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A Survey of Hallucination in Large Foundation Models. [cs.AI] <https://arxiv.org/abs/2309.05922>
- [142] Patrik Reizinger and Wieland Brendel. 2026. HALLMARK: Diagnosing Three Failure Modes in LLM Citation Verifiers. [cs.CR] <https://arxiv.org/abs/2607.18360>
- [143] Baochang Ren, Shuofei Qiao, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025. KnowRL: Exploring Knowledgeable Reinforcement Learning for Factuality. [cs.AI] <https://arxiv.org/abs/2506.19807>
- [144] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4035–4045. [cs.CL] doi:10.18653/v1/D18-1437
- [145] Pranab Sahoo, Prabhaskar Mehar, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 11709–11724. [cs.LG] doi:10.18653/v1/2024.findings-emnlp.685
- [146] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. [cs.CL] <https://arxiv.org/abs/2402.03300>
- [147] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. [cs.CL] <https://arxiv.org/abs/2310.13548>
- [148] Xu Shen, Song Wang, Zhen Tan, Laura Yao, Xinyu Zhao, Kaidi Xu, Xin Wang, and Tianlong Chen. 2026. FaithCoT-Bench: Benchmarking Instance-Level Faithfulness of Chain-of-Thought Reasoning. In *International Conference on*

- Learning Representations*. OpenReview.net, Rio de Janeiro, Brazil, 31 pages. [cs.AI] <https://openreview.net/forum?id=IN3yKqzF1>
- [149] Jerick Shi, Terry Jingcheng Zhang, Zhijing Jin, and Vincent Conitzer. 2026. From Hallucination to Scheming: A Unified Taxonomy and Benchmark Analysis for LLM Deception. arXiv v1, 6 April 2026. [cs.CY] <https://arxiv.org/abs/2604.04788v1>
 - [150] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. [cs.CL] <https://arxiv.org/abs/2305.14739>
 - [151] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), New Orleans, LA, USA, 8634–8652. [cs.AI] doi:10.52202/075280-0377
 - [152] Shuzheng Si, Qingyi Wang, Haozhe Zhao, Yuzhuo Bai, Guanqiao Chen, Kangyang Luo, Gang Chen, Fanchao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. 2025. FaithLens: Detecting and Explaining Faithfulness Hallucination. [cs.CL] <https://arxiv.org/abs/2512.20182>
 - [153] Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025. The Hallucination Tax of Reinforcement Finetuning. [cs.CL] <https://arxiv.org/abs/2505.13988>
 - [154] Peiyang Song, Pengrui Han, and Noah Goodman. 2026. Large Language Model Reasoning Failures. Transactions on Machine Learning Research. Survey Certification, January 2026. [cs.AI] <https://arxiv.org/abs/2602.06176>
 - [155] Zhongxiang Sun, Qipeng Wang, Haoyu Wang, Xiao Zhang, and Jun Xu. 2025. Detection and Mitigation of Hallucination in Large Reasoning Models: A Mechanistic Perspective. [cs.AI] <https://arxiv.org/abs/2505.12886>
 - [156] Hamed Taheri. 2026. Governed Memory: A Production Architecture for Multi-Agent Workflows. <https://arxiv.org/abs/2603.17787v1>
 - [157] Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. 2025. Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Association for Computational Linguistics, Suzhou (China), 799–811. doi:10.18653/v1/2025.emnlp-industry.54
 - [158] Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. [cs.CL] <https://arxiv.org/abs/2404.10774>
 - [159] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. [cs.CL] <https://arxiv.org/abs/2207.04672>
 - [160] Yuchen Tian, Weixiang Yan, Qian Yang, Xuandong Zhao, Qian Chen, Wen Wang, Ziyang Luo, Lei Ma, and Dawn Song. 2024. CodeHalu: Investigating Code Hallucinations in LLMs via Execution-based Verification. [cs.CL] <https://arxiv.org/abs/2405.00253>
 - [161] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. [cs.CL] <https://arxiv.org/abs/2302.13971>
 - [162] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udel, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering Language Models With Activation Engineering. [cs.CL] <https://arxiv.org/abs/2308.10248>
 - [163] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. [cs.CL] <https://arxiv.org/abs/2305.04388>
 - [164] Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 9935–9960. [cs.CL] doi:10.18653/v1/2025.emnlp-main.504
 - [165] Sanidhya Vijayvargiya and Rahul Lokes. 2026. Actionable Hallucination Detection: Translating Latent Uncertainty into Agentic Critique. [cs.LG] <https://arxiv.org/abs/2608.10430>
 - [166] Changyue Wang, Weihang Su, Qingyao Ai, and Yiqun Liu. 2025. Joint Evaluation of Answer and Reasoning Consistency for Hallucination Detection in Large Reasoning Models. [cs.CL] <https://arxiv.org/abs/2506.04832>
 - [167] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2023. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. [cs.CL] <https://arxiv.org/abs/2311.07397>
 - [168] Kaishen Wang, Tong Zheng, Xuehao Cui, Ruibo Chen, Tianyi Xiong, and Heng Huang. 2026. Mitigating Factual Hallucination in Large Reasoning Models via Mixed-Mode Advantage Regularization. [cs.CL] <https://arxiv.org/abs/>

- 2607.05861
- [169] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. [cs.CL] <https://arxiv.org/abs/2203.11171>
 - [170] Yanbo Wang, Yongcan Yu, Jian Liang, and Ran He. 2025. A Comprehensive Survey on Trustworthiness in Reasoning with Large Language Models. [cs.CL] <https://arxiv.org/abs/2509.03871>
 - [171] Ziming Wang, Yinghua Yao, Changwu Huang, Ke Tang, and Xin Yao. 2026. CASE: Causal Alignment and Structural Enforcement for Improving Chain-of-Thought Faithfulness. [cs.CL] <https://arxiv.org/abs/2607.18820>
 - [172] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. [cs.CL] <https://arxiv.org/abs/2411.04368>
 - [173] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35*. Neural Information Processing Systems Foundation, Inc. (NeurIPS), New Orleans, LA, USA, 24824–24837. [cs.CL] doi:10.52202/068431-1800
 - [174] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. [cs.CL] <https://arxiv.org/abs/2403.18802>
 - [175] Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs. [cs.AI] <https://arxiv.org/abs/2506.14245>
 - [176] Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. 2026. LongMemEval-V2: Evaluating Long-Term Agent Memory Toward Experienced Colleagues. <https://arxiv.org/abs/2605.12493v1>
 - [177] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=pZiyCaVuti>
 - [178] Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. 2025. Mitigating LLM Hallucination via Behaviorally Calibrated Reinforcement Learning. [cs.LG] <https://arxiv.org/abs/2512.19920>
 - [179] Yuhe Wu, Guangyu Wang, Yuran Chen, Jiatong Zhang, Yutong Zhang, Yujie Chen, Jiaming Shang, Guang Zhang, and Zhuang Liu. 2026. PRISM: Probing Reasoning, Instruction, and Source Memory in LLM Hallucinations. [cs.CL] <https://arxiv.org/abs/2604.16909>
 - [180] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. [cs.CL] <https://arxiv.org/abs/2306.13063>
 - [181] Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 13326–13365. doi:10.18653/v1/2024.acl-long.720
 - [182] Mengzhu Xu, Jifan Gao, Xia Jiang, Yaoxin Wu, and Xi Long. 2026. Right Diagnoses, Decorative Reasoning: A Perturbation Audit of Medical Chain-of-Thought. [cs.AI] <https://arxiv.org/abs/2608.24790>
 - [183] Sen Xu, Wei Wang, Shixi Liu, Jixin Min, Yingwei Dai, Zhibin Yin, Yirong Chen, and Junlin Zhang. 2026. Claim-Level Reliability Assessment for Efficient Test-Time Reasoning. [cs.AI] <https://arxiv.org/abs/2608.11994>
 - [184] Cheng Yan, Zhijun Fan, Guangyang Ye, Fan Xu, Xiang Xia, Yawei Wang, and Wuyang Zhang. 2026. UPAIR: Diagnosing Reasoning States via Uncertainty-Progress Alignment for Selective Intervention. [cs.AI] <https://arxiv.org/abs/2607.17188>
 - [185] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. [cs.CL] <https://arxiv.org/abs/2505.09388>
 - [186] Tianyu Yang, Sudipta Paul, Vijay Srinivasan, Vivek Kulkarni, and Srinivas Chappidi. 2026. TRUSTMEM: Learning Trustworthy Memory Consolidation for LLM Agents with Long-Term Memory. <https://arxiv.org/abs/2606.25161v1>
 - [187] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. [cs.CL] <https://arxiv.org/abs/2210.03629>

- [188] Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. 2025. Are Reasoning Models More Prone to Hallucination? [cs.CL] <https://arxiv.org/abs/2505.23646>
- [189] Chenlong Yin, Zeyang Sha, Shiwen Cui, Changhua Meng, and Zechao Li. 2025. The Reasoning Trap: How Enhancing LLM Reasoning Amplifies Tool Hallucination. [cs.LG] <https://arxiv.org/abs/2510.22977>
- [190] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. [cs.LG] <https://arxiv.org/abs/2503.14476>
- [191] Xiangning Yu, Zhuohan Wang, Linyi Yang, Haoxuan Li, Anjie Liu, Xiao Xue, Jun Wang, and Mengyue Yang. 2025. Causal Sufficiency and Necessity Improves Chain-of-Thought Reasoning. [cs.CL] <https://arxiv.org/abs/2506.09853>
- [192] Zhe Yu, Wenpeng Xing, Yunzhao Wei, Jie Chen, Hongzhi Wang, Xuyang Teng, and Meng Han. 2026. Composition Collapse: Stable Factual Knowledge Does Not Imply Compositional Reasoning. [cs.AI] <https://arxiv.org/abs/2605.26789>
- [193] Zhe Yu, Wenpeng Xing, Chen Ye, Xuyang Teng, Bo Yang, Changting Lin, and Meng Han. 2026. Detecting Is Not Resolving: The Monitoring Control Gap in Retrieval Augmented LLMs. [cs.AI] <https://arxiv.org/abs/2605.27157>
- [194] Jennifer Za, Julija Bainiaksina, Nikita Ostrovsky, Tanush Chopra, and Victoria Krakovna. 2026. Persuasion Attacks Can Decrease Effectiveness of CoT Monitoring. [cs.AI] <https://arxiv.org/abs/2607.08066>
- [195] Skylar Zhai, Jingcheng Liang, and Dongyeop Kang. 2026. Abstain-R1: Calibrated Abstention and Post-Refusal Clarification via Verifiable RL. [cs.CL] <https://arxiv.org/abs/2604.17073>
- [196] Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. 2026. Adaptive Memory Admission Control for LLM Agents. <https://arxiv.org/abs/2603.04549v1>
- [197] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics* 51, 4 (2025), 1373–1418. doi:10.1162/coli.a.16
- [198] James Xu Zhao, Bryan Hooi, and See-Kiong Ng. 2025. Test-Time Scaling in Reasoning Models Is Not Effective for Knowledge-Intensive Tasks Yet. [cs.AI] <https://arxiv.org/abs/2509.06861>
- [199] Xiangyu Zhao, Henry Hengyuan Zhao, Yiheng Wang, Wanghan Xu, Yuhao Zhou, Qinglong Cao, Zhiwang Zhou, Lei Bai, Wenlong Zhang, and Xiao-Ming Wu. 2026. Sci-PRM: A Tool Aware Process Reward Model for Scientific Reasoning Verification. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*. ACM, New York, NY, USA, 13008–13019. [cs.AI] doi:10.1145/3770855.3818999
- [200] Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. [cs.CL] <https://arxiv.org/abs/2305.14795>
- [201] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. [cs.AI] <https://arxiv.org/abs/2205.10625>
- [202] Wei Zhou, Xuanhe Zhou, Shaokun Han, Hongming Xu, Guoliang Li, Zhiyu Li, Feiyu Xiong, and Fan Wu. 2026. Are We Ready For An Agent-Native Memory System? <https://arxiv.org/abs/2606.24775v1>
- [203] Yifan Zhu, Huiqiang Rong, and Haoran Luo. 2026. Token-Guard: Towards Token-Level Hallucination Control via Self-Checking Decoding. [cs.CL] <https://arxiv.org/abs/2601.21969>
- [204] Geigh Zollicoffer, Minh Vu, Hongli Zhan, Raymond Li, and Manish Bhattarai. 2026. Sanity Checks for Long-Form Hallucination Detection. [cs.CL] <https://arxiv.org/abs/2605.08346>
- [205] Ádám Kovács, Bowei He, Xue Liu, István Boros, Szilveszter Tóth, and Gábor Recski. 2026. Beyond Document Grounding: Span-Level Hallucination Detection over Code, Tool Output, and Documents. [cs.CL] <https://arxiv.org/abs/2607.00895>