WHAT HALLUCINATIONS HAVE WE MEASURED? AN EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains one of the most critical challenges in deploying large language models (LLMs) reliably and safely. Yet the term is used inconsistently across the research community. Analyzing hallucination-titled papers from major NLP and ML venues, we find that explicit definitions are increasingly rare and that the term conflates fundamentally different error types while failing to distinguish hallucinations from *expressed uncertainty*[1]. This ambiguity obscures what is being measured and why models fail, limiting progress in detection and mitigation. In this work, we show that hallucinated responses should be distinguished from expressed uncertainty and that hallucination is not a unitary phenomenon. We identify two types of model incorrect responses: *confabulation*, where a model consistently produces incorrect responses, and *semantic illusion*, where the response correctness varies with the prompt. These two types exhibit distinct empirical signatures and are systematically conflated in existing evaluations. We formalize this distinction with operational definitions and diagnostic criteria, and show that separating them improves interpretability and reveals limitations of current mitigation methods. We advocate for future work to explicitly distinguish and report these error types.

1 INTRODUCTION

Hallucinations in large language models (LLMs) have emerged as a central concern for both researchers and practitioners. Yet there is surprisingly little consensus on what the term actually means. Across the literature, it is applied to a broad and heterogeneous set of model failures—factual inaccuracies [25, 2, 1], faithfulness errors [13, 34], reasoning inconsistencies [12, 6], and uncertain outputs [34, 12, 31]—often without distinction or explicit definition. This conceptual ambiguity is not merely terminological but obscures the underlying causes of model incorrect responses, undermines evaluation validity, and hinders the development of targeted mitigation strategies.

Recent hallucination-titled papers reveal two systematic methodological gaps (Fig. 1). The first is the hallucination-uncertainty[2] distinction in definitions and methodologies. Only 35% of papers nominally define hallucination as confident or plausible errors rather than *expressed uncertainty*, and even this is not always reflected in methodology. While 23% treat hallucination detection as distinct from *uncertainty estimation*[3], not all anchor this in a confidence-based definition. Researchers broadly agree this distinction matters, as confirmed by a survey of 24 first authors of hallucination papers in top ML/NLP venues. 95.8% believe hallucination should be distinguished from uncertainty (both expressed and estimated), yet only 16.7% report treating them as distinct in their work. We address this mismatch, a gap whose practical consequences we define in § 3.2 and quantify in § 5.3.

The second gap is knowledge diagnosis. Among papers engaging with hallucination and uncertainty, those concluding that new external knowledge is necessary (15% of total) mostly never tested prompt variance first (11% of total), despite many incorrect responses yielding correct ones under different question phrasings, instruction prompts, or inference-time decoding configurations (§ 2 and

[1] Model responses that explicitly communicate epistemic uncertainty, including both abstention (i.e., declining to provide an answer) and low-confidence answers (e.g., "I am not sure but maybe ...", "I don't know but ...").

[2] Note that uncertainty here encompasses both *expressed uncertainty* and *uncertainty estimation* methods, which are treated as distinct concepts throughout the paper.

[3] Methods that quantify epistemic uncertainty in model outputs, e.g., via token probabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

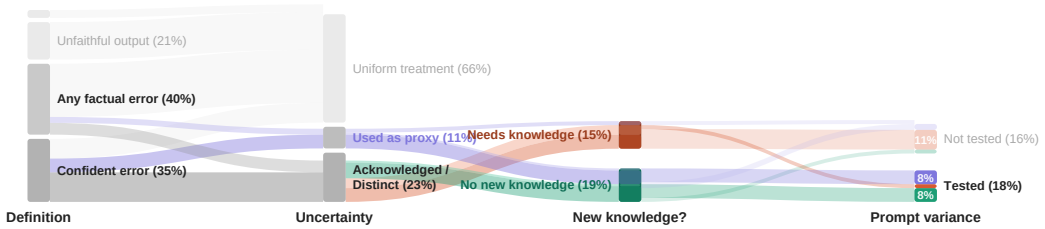


Figure 1: **Recent text-based hallucination papers filtered by their definitions and approaches with uncertainty estimation, additional knowledge, and prompt variance.** From a pool of 77 qualifying text-based papers across ICML, NeurIPS, ICLR, ACL, and EMNLP (2023–2025), most papers neither distinguish hallucination from uncertainty nor test prompt variance before considering external knowledge as a source. The methods that skip these distinctions risk targeting the wrong error type entirely.

*We did not filter papers further if they use the term "hallucination" in the title and take a text-based approach.

§ 3.1). Among those that do not introduce new knowledge (19% of total), some never tested prompt variance either (3% of total papers), potentially leaving their results suboptimal when simple prompt interventions could yield further gains. Without these diagnostics, methods risk over-engineering solutions to errors that simpler interventions would resolve. This concern is crucial in hallucination research with 79.2% of surveyed researchers identifying knowledge availability, whether a model lacks knowledge versus misapplies it, as the most critical dimension for characterizing hallucinations (Fig. 9).

To address this issue, we empirically study 14 recently published works on hallucinations from major NLP and ML venues and argue that two distinctions are essential but consistently overlooked. First, hallucinations should be separated from expressed uncertainty (UNK). Second, hallucinations are not homogeneous but differ in behavior patterns that point to the underlying cause. We distinguish two subtypes: (i) *confabulation* (CF), errors that persist across all plausible evaluation conditions such as prompt phrasings, instruction variants, and inference-time decoding configurations; and (ii) *semantic illusion* (SI), errors that vary with these same conditions. We take prompt-invariance as an operational proxy for knowledge availability; this is not a claim about internal knowledge states. This predicts that CF should require external knowledge whereas SI should not, which we observe in § 5.2. The practical implication is to exhaust the inexpensive inference-time or post-hoc interventions before turning to costly retrieval or retraining.

We establish three empirical findings. First, detection methods assign higher hallucination rates to UNK, where the model explicitly expresses uncertainty, than to confidently incorrect responses (SI and CF). Second, the mitigation gap between CF and SI is severe, as flagged SI responses are corrected 28.5% of the time and CF responses 0.7%. Third, uncertainty estimation calibrated on output-level signals detects UNK reliably but degrades on CF and SI, since confidently incorrect responses carry no output-level uncertainty cue; latent representations, which bypass surface signals, partially close this gap. These findings collectively demonstrate that a single hallucination score is insufficient and that progress requires methods sensitive to the behavioral distinctions between these error types.

2 Background

Definition Despite its widespread usage, the term hallucination lacks a consensus definition across the research community. The proportion of papers explicitly defining it has declined steadily since 2021 [24]. Early NLP usage anchored the term to faithfulness failures, describing generated content unfaithful to a provided source [13, 34]. With the rise of LLMs, definitions have diverged along two axes. The first concerns the nature of ***incorrectness***: from source unfaithfulness [13, 34], to real-world unfactualness covering fictional, misleading, or fabricated outputs [25, 2, 1], and further to outputs conflicting with instructions, context, or logical consistency [6, 12]. The second concerns the degree of ***plausibility***: from outputs that simply appear fluent and coherent [1, 2], to those whose plausibility is treated as an unavoidable consequence of how LLMs generate text [33, 17], to the strictest view requiring misplaced confidence, restricting hallucination to assertions the model delivers with unwarranted certainty [15, 24, 12].

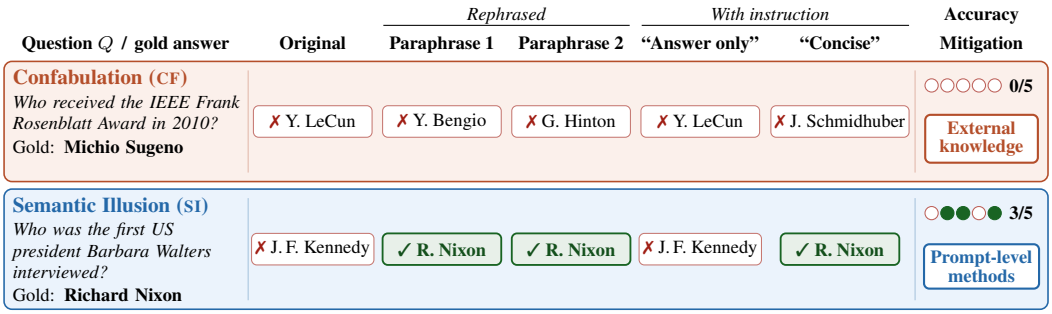


Figure 2: **Worked example contrasting CF and SI.** Both questions yield an incorrect response on the original prompt; the response pattern across semantically equivalent phrasings and instruction variants determines the behavioral category and predicts which intervention applies.

Terminology The terminological ambiguity has prompted calls for more precise alternatives. A systematic study of researcher perspectives [24] found that 40.46% prefer *fabrication*, while others advocate for *confabulation*, a clinical term for false narratives produced to bridge knowledge gaps without intent to deceive [3, 22, 26]. This preference reflects a broader concern that models frequently fabricate responses without sufficient grounds. The clinical origin of the term, however, is more specific, denoting errors arising from missing or inaccessible knowledge [3, 22, 26, 8, 20]. Under this view, responses may vary in form across semantically equivalent prompt variants, yet remain incorrect regardless of how the question is phrased. This captures only one error type. We therefore retain *hallucination* as the cover term and decompose it into two behaviorally distinct subtypes, *confabulation* and *semantic illusion*, whose empirically separable signatures we define in § 3.1 and analyze in § 5.

Taxonomy Early survey work established the foundational intrinsic/extrinsic taxonomy, where intrinsic hallucinations contradict the source content and extrinsic hallucinations introduce unverifiable information [13]. Subsequent work expanded this line of taxonomies. [34] proposed input-conflicting, context-conflicting, and fact-conflicting categories to capture misaligned responses beyond source documents, while [12, 29] separated factuality hallucination from faithfulness hallucination across a hierarchical set of subtypes. [2] attributed hallucinations to either prompting or model-internal factors by measuring how errors change across prompt variants and across models. Unlike prior taxonomies based on triggers or contradictions, we classify hallucinations by prompt sensitivity, distinguishing consistent incorrectness (*confabulation*) from prompt-sensitive incorrectness (*semantic illusion*), and linking each to corresponding detection and mitigation strategies.

3 Problem Formulation

We formalize the two central hallucination dimensions – *incorrectness* and *plausibility* – through the following research questions (RQ) and hypotheses (H).

3.1 RQ1 (Incorrectness): Do hallucinations arise from distinct error types?

The most basic condition for hallucinations is incorrectness, whether measured against a source [13, 34], the real-world facts [25, 2, 1], or an instruction [12, 6]. Incorrect outputs, however, are not uniform in how they arise or how they respond to elicitation. Accordingly, the difference that matters most in practice is consistency of occurrence. Yet some incorrect responses can be corrected under a semantically equivalent phrasing (*semantic illusion*), while others recur whatever phrasing is used (*confabulation*). We ask whether these reflect distinct error types, and whether that distinction changes which detection and mitigation strategies apply.

The difference also invites an epistemic reading, since a response that becomes correct under a semantically equivalent phrasing shows the relevant knowledge is accessible without changing model weights or injecting additional information. We therefore treat knowledge as behaviorally accessible when some tested condition elicits a correct response. The conditions we admit that establish accessibility are semantically equivalent phrasings, instruction variants, decoding configurations (e.g.,

temperature, nucleus sampling), which require no access to model internals and add no additional hints or context.

Notation. Let $\mathcal{V}(Q)$ be a set of semantically equivalent variants of question Q sharing the same ground truth but introducing no additional hints. For each variant $v \in \mathcal{V}(Q)$, $r(v)$ denotes the model response; $\text{correct}(r(v)) = 1$ if it matches the ground truth, $\text{abstain}(r(v)) = 1$ if the model expresses uncertainty (confidence score $= 0$ or explicit refusal).

Definition 1 Confabulation (CF)

A response to question Q is a **confabulation** if the model produces an incorrect, non-abstaining answer on *every* variant in $\mathcal{V}(Q)$ under all plausible inference-time conditions (e.g., instruction variants, decoding configurations):

$$\text{CF}(Q) \iff \forall v \in \mathcal{V}(Q) : \neg\,\text{correct}(r(v)) \wedge \neg\,\text{abstain}(r(v)).$$

Definition 2 Semantic Illusion (SI)

A response to question Q is a **semantic illusion** if the model produces an incorrect answer on at least one variant and a correct answer on at least one other variant, without any change to model parameters or the introduction of additional information:

$$\text{SI}(Q) \iff \exists v_1, v_2 \in \mathcal{V}(Q) : \neg\,\text{correct}(r(v_1)) \wedge \text{correct}(r(v_2)).$$

Remark 1 (Stability relative to $\mathcal{V}(Q)$). CF and SI are defined with respect to a finite evaluation set $\mathcal{V}(Q)$. Expanding $\mathcal{V}(Q)$ may reclassify some CF instances as SI if a correct response is elicited. We therefore treat them as *operational diagnostics*, not intrinsic model properties. This reflects practical constraints. Model behavior can only be probed under a limited set of prompts and decoding settings. Under this constraint, CF corresponds to cases where no tested condition yields a correct response, while SI corresponds to errors recoverable within the same evaluation budget. Although labels may change as $\mathcal{V}(Q)$ grows, the mitigation gap between CF and SI remains consistent across schemes. We show the robust empirical observations despite different $\mathcal{V}(Q)$ in § 4.1.

H1 (Detection asymmetry). If confabulation (CF) and semantic illusion (SI) reflect genuinely distinct error types, existing detection methods should perform asymmetrically across them. Because many hallucination methods rely on inconsistency signals, they will detect SI more reliably than CF.

H2 (Intervention asymmetry). CF and SI should respond differently to mitigation. Inference-time and post-hoc interventions will reduce SI errors but leave CF errors largely unchanged, since CF reflects behaviorally inaccessible knowledge that prompt manipulation cannot mitigate. Mitigating CF should depend on external knowledge.

3.2 RQ2 (Plausibility): Does hallucination require confident incorrectness?

We define plausibility as confidence, because (i) current LLMs are fluent by default, so surface plausibility is satisfied by nearly every output and draws no useful boundary, and (ii) a confidently stated incorrect response gives the reader no cue to check it, which makes it the more dangerous error. We say a response is **confident** if it does not explicitly express uncertainty. A confidently incorrect response (CF, SI) and an expressed uncertainty (UNK) may both be scored as incorrect, but reflect qualitatively different behaviors and risks. Uncertainty estimation methods were originally developed to quantify epistemic uncertainty as a signal for detecting incorrect outputs, whereas UNK is a *behavior* in which the model explicitly abstains. As LLMs increasingly exhibit this behavior, the remaining errors are predominantly confident ones, where output-level uncertainty signals have least to detect. We ask whether confidence is a necessary condition for hallucination and whether existing methods respect this distinction in practice.

Definition 3 Expressed Uncertainty (UNK)

A response to question Q is an **expressed uncertainty** if the model produces an incorrect answer on every variant in $\mathcal{V}(Q)$ but abstains on at least one.

Table 1: **Experimental dimensions for RQ1–RQ2.** H1: prompt sensitivity, intervention stage, and target error type; H2: knowledge assumption; H3–H4: uncertainty usage and representation. Dimensions are explained in § B.

RQ	H	Angle	Category	Methods
RQ1	H1	Prompt Sensitivity	Modeled	SECA [18], SelfCheckGPT [19]
			Not modeled	DRAG [7], UQ Heads [27], ActLCD [32]
RQ1	H1	Intervention Stage	Inference	ActLCD [32], NoVo [10]
			Post-hoc	SelfCheckGPT [19], UQ Heads [27]
			Ext. Know.	DRAG [7], HEB [11], CS(Beyond Fact) [9]
RQ1	H2	Knowledge Assump.	Absent	SelfCheckGPT [19], IRIS [28], VUF [14], SCH [23], EES(SenD) [21], Attn. Sink [5]
			Present	NoVo [10], ActLCD [32], LapEigvals [4], SECA [18]
RQ2	H3	Uncertainty Usage	Calibration	VUF [14], UQ Heads [27], IRIS [28], Attn. Sink [5], SelfCheckGPT [19]
			Correction	EES(SenD) [21], SCH [23]
RQ2	H4	Uncertainty Repr.	Verbal	VUF [14], UQ Heads [27]
			Latent	UQ Heads [27], EES(SenD) [21], IRIS [28], Attn. Sink [5]
			Semantic	SelfCheckGPT [19], SCH [23], IRIS [28]
			None	SECA [18], DRAG [7], HEB [11], NoVo [10], ActLCD [32], LapEigvals [4]

H3 (Uncertainty asymmetry). If detection methods conflate UNK with hallucination, they will assign higher hallucination scores to UNK than to CF and SI, flagging the less dangerous error type more readily than the more dangerous ones.

H4 (Representation asymmetry). Among uncertainty-based methods, verbal representations will be most effective at detecting UNK, which carries an explicit uncertainty cue in its output. In contrast, CF carries no such cue. Therefore, latent representations, which read internal states beyond the output, will be most effective at detecting and mitigating CF.

4 METHODS

Study Design We analyze recent hallucination research and conduct a controlled empirical evaluation. We collect 14 papers (2023–2025) from major ML and NLP venues with "hallucination" in the title, and categorize them based on methodological choices reported by the authors. Dimensions are organized around our two research questions, as summarized in Table 1 and visualized in Figure 3. For RQ1, we examine prompt sensitivity (whether prompt variation is modeled), intervention stage (inference-time vs. post-hoc vs. external knowledge), and model knowledge assumption (whether the method implicitly assumes knowledge is absent or present). For RQ2, we examine the usage of uncertainty estimation (whether uncertainty is used for calibration or correction) and uncertainty representation (verbal, latent, semantic, or none). Together, these dimensions position existing methods within our taxonomy and show how their design choices relate to different error types.

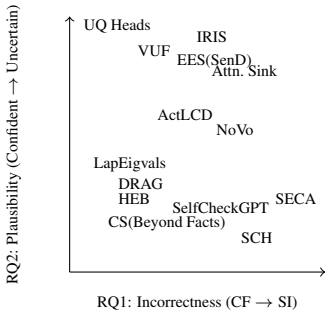


Figure 3: **Methods in RQ1–RQ2 space.**

Dataset We construct a controlled factual QA dataset drawn from TriviaQA [16] and SimpleQA [30] across (NH,SI,CF,UNK) categories for our study. We study closed-book factual question answering, because it provides verifiable ground truth, makes model knowledge the sole explanatory variable, and removes retrieval or grounding confounds. We assign them by 30 semantically equivalent question variants for a practical evaluation. We report all main results across two different instruction methods (see details in § C). **S1** applies a richer grid of 30 semantically equivalent variants, 3 instruction prompts without uncertainty elicitation, and multiple inference-time decoding configurations (e.g., temperatures, see details in § C.3); **S2** additionally includes an uncertainty-elicitation instruction.

Table 2: Label distribution

Scheme	CF	SI	UNK
S1	0.36	0.62	0.02
S2	0.04	0.64	0.32

Evaluation We evaluate detection and mitigation on Llama 3.1 8B and Gemma 3 12B, grouping methods by shared design dimensions (Table 1) and reporting mean ± std across methods. Let $h \in \{0, 1\}$ indicate hallucination prediction, $c \in \{0, 1\}$ correctness, and $\ell \in \{\text{SI}, \text{CF}, \text{UNK}\}$ error type; incorrect responses (c=0) are treated as hallucinations. We report detection and mitigation metrics stratified by ℓ. We report *recall* $P(h{=}1 \mid c{=}0, \ell)$, *residual error rate* $P(c{=}0 \mid h{=}0, \ell)$, and *mitigation correctness* $P(c{=}1 \mid h{=}1, \ell)$.

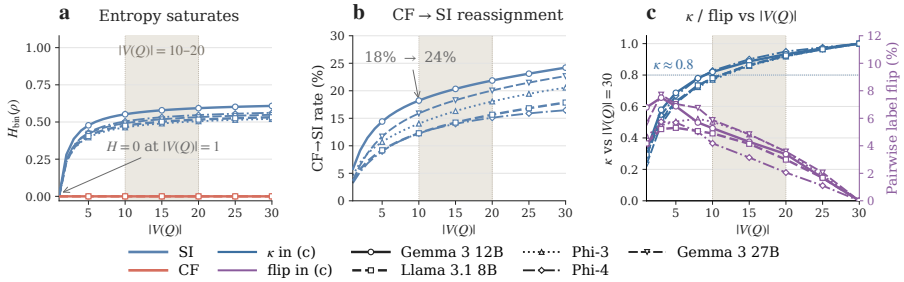


Figure 4: **CF/SI label behavior as $|\mathcal{V}(Q)|$ increases, subsampled from a 30-paraphrase pool.** (a) binary entropy of the correctness rate. (b) CF → SI reassignment rate. (c) Cohen's κ against the full-pool label and pairwise subset flip rate. Shading marks the recommended window $|\mathcal{V}(Q)|$.

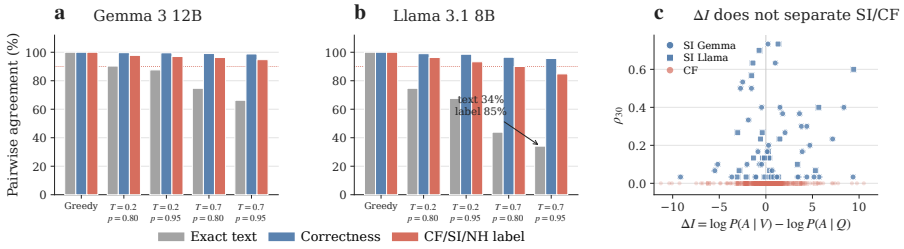


Figure 5: **Label stability under decoding, and SI/CF is not an information-injection artifact.** (a)&(b) Pairwise agreement (%) on a 30 question variants, aggregated over three instruction prompts. (c) ρ_{30} vs. the information quantity emerged from paraphrases $\Delta I = \log P(A \mid V) - \log P(A \mid Q)$.

4.1 Is Accessible Knowledge Measurable? Behavioral Evidence from $\mathcal{V}(Q)$

We define knowledge as behaviorally *accessible* if at least one variant $v \in \mathcal{V}(Q)$ elicits a correct response without additional hints or parameter changes. Note that the $\mathcal{V}(Q)$ is used only for labeling and is never exposed to evaluated methods. This section establishes three properties of that criterion; (i) the labels stabilize with a modest number of paraphrases, (ii) the correct answers under semantically equivalent phrasings reflect accessible knowledge rather than added information, and (iii) the label convergence can verify whether $|\mathcal{V}(Q)|$ is sufficient.

Sufficiency of $\mathcal{V}(Q)$ A practitioner can judge whether $\mathcal{V}(Q)$ is large enough by tracking label agreement as it grows. Let ρ be the fraction of paraphrases answered correctly, conditioned on the original question being wrong. We subsample $|\mathcal{V}(Q)|$ variants from a 30 paraphrase pool. We then track how labels change and show how we can verify the sufficiency of $\mathcal{V}(Q)$ (Fig. 4). First, the SI entropy plateaus at certain $|\mathcal{V}(Q)|$. Second, the labels reassignment from CF to SI flattens. Third, the agreement with full number of labels reaches high κ scores. Adding paraphrases can move a question CF→SI but never the reverse, since adding more paraphrases cannot revoke already correct answer incorrect. Therefore, the reported CF rate may provide upper bound of inaccessible knowledge but almost impossible to reach the lower bound.

Correct answers under paraphrasing reflects accessibility, not added information. SI shows relatively low but nonzero ρ (median $\rho_{30} = 0.17$ / 0.13 by Gemma / Llama), whereas CF shows none ($\rho_{30} = 0$) (Fig. 5c). This leads H_{bin} of SI to saturate near half a bit rather than 1 (Fig. 4a). This means that the correct answer is reachable under some rewording but not reliably elicited by it. To rule out an unintended hint, we report the information quantity of each variant V relative to the original question Q, $\Delta I = \log P(A \mid V) - \log P(A \mid Q)$, with A the gold answer (Fig. 5c). ΔI overlaps around zero for both SI and CF, though only SI gets correct answer, so this mitigation cannot be attributed to a phrasing more informative than the original question.

Recommended reliability protocol. (1) Use approximately 10 variants ($|\mathcal{V}(Q)| \approx 10$) as a default and label under greedy or low-temperature decoding. Because SI and CF labels remain largely stable across decoding settings even when the exact response text changes (Fig. 5a,b), we take this setup as a practical approximation of the full labeling conditions. (2) Report the H_{bin} (and κ / flip-rate) convergence curve on one's own data as a sufficiency check. (3) Treat CF as an upper-bound on

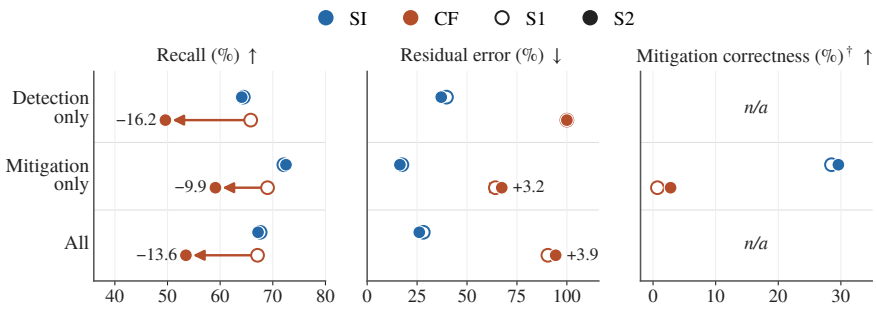


Figure 6: **SI vs. CF across labeling schemes at $|\mathcal{V}(Q)| = 10$.** Each marker is the mean over methods in that group. Mitigation correctness, $^{\dagger}P(c{=}1 \mid h{=}1, \ell)$, is for mitigation methods only.

behavioral inaccessibility. Additional variants can only move CF error cases to SI ones. These steps makes the taxonomy reproducible across studies with a modest number of standard paraphrases.

5 Results

Across all methodological dimensions, models, and labeling schemes, one asymmetry is consistent: SI is both detectable and mitigable, whereas CF is rarely mitigated and, under S2, less often detected. In this section, we empirically study that this is a property of CF itself, rather than of any individual method.

5.1 The CF Blind Spot

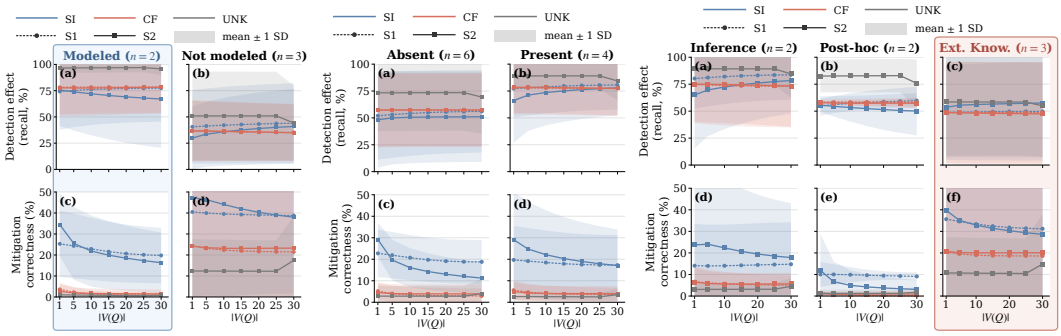
SI is detectable and mitigable but CF is less detectable and hard to mitigate. Figure 6 compares SI and CF under S1 and S2, the two labeling schemes without and with uncertainty elicitation. The SI and CF differ in both detection (recall and residual error) and mitigation (mitigation correctness). The detection is asymmetric, as SI scores generally better than CF. The CF responses become harder to detect as the labeling scheme changes from S1 to S2, as shown in recall scores (red). One possible explanation is that because the uncertainty-elicitation prompt moves the easier CF responses into UNK, the CF set under S2 retains only the harder questions (see the next paragraph). The mitigation shows even sharper asymmetric patterns. The mitigation correctness reaches 28.5% and 29.6% for SI under S1 and S2 respectively, but only 0.7% and 2.8% for CF. This means that even when CF responses are detected, they are rarely mitigated while SI is both detected and mitigated at a meaningful rate.

Uncertainty elicitation migrates CF to UNK, rather than correcting it. Adding an uncertainty-elicitation instruction under S2 moves most CF responses into UNK. The CF proportion falls from 0.36 to 0.04 while UNK rises from 0.02 to 0.32 (Table 2). The CF set under S2 is responses that stay confidently incorrect even when uncertainty elicitation is offered. Yet this relabeling does not make CF any easier to correct (Fig. 6 mitigation correctness).

5.2 Intervention Asymmetry

Prompt sensitivity methods detect CF on par with SI but mitigate SI only, performing the strongest SI mitigation among all method dimensions (Fig. 7a) . They detect CF and UNK well but mitigate them at near zero rate. Notably, their SI detection under S2 decreases from approximately 76% to 67% as $|\mathcal{V}(Q)|$ increases, the only notable decline in detection across the method dimensions. A possible explanation is that larger $|\mathcal{V}(Q)|$ filters out SI errors readily exposed by rephrasing, leaving more difficult errors that prompt sensitivity methods struggle to detect. Methods that do not model prompt sensitivity show the complementary pattern, with weaker detection but better CF and UNK mitigation. This suggests a potential composition in which prompt sensitivity identifies errors and stronger corrective methods that are outside the prompt levels mitigate the detected cases.

Methods assuming knowledge is present (Present methods) outperforms those assuming absent (Absent methods) in detection for all three types (SI, CF, and UNK) and mitigation for hard SI under S2 (Fig. 7b). Present methods also exhibit increasing SI detection under S2 as $|\mathcal{V}(Q)|$ grows. This detection advantage is noteworthy because S2 labeling scheme is likely to remain difficult SI

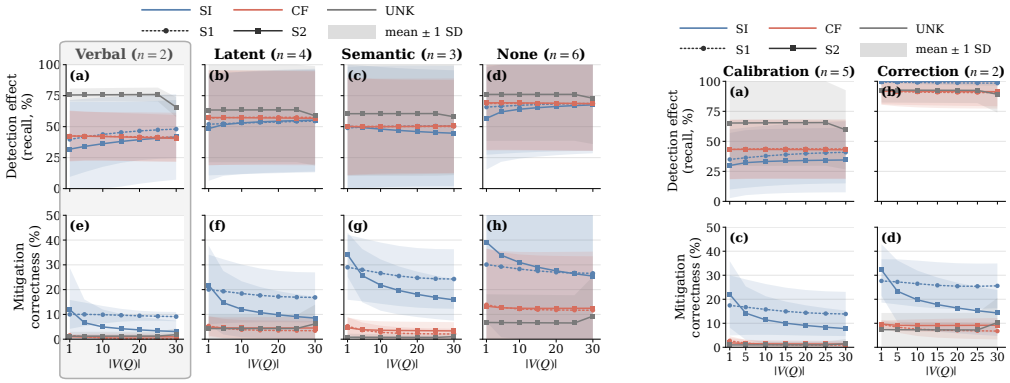


(a) Prompt Sensitivity (b) Knowledge Assumption (c) Intervention Stage

Figure 7: **Comparison of existing methods.**

Table 3: **Uncertainty estimation-based detection and mitigation performance** (mean±std).

	Detection						Mitigation					
	S1			S2			S1			S2		
Metric	CF	SI	UNK	CF	SI	UNK	CF	SI	UNK	CF	SI	UNK
Miss rate (%) ↓	34.2±25.6	35.6±24.7	**22.9±21.4**	50.4±33.5	35.9±24.9	**30.8±23.0**	–	27.9±37.6	–	41.1±44.7	**28.1±37.7**	29.6±38.1
Hallucination score ↑	0.54±0.21	0.43±0.20	**0.61±0.25**	0.42±0.29	0.42±0.21	**0.57±0.19**	–	–	–	–	–	–
Correctness (%) ↑	–	–	–	–	–	–	3.9±7.3	**29.5±7.5**	0.0±0.0	2.2±5.0	**30.4±8.6**	0.4±0.7



(a) Uncertainty representation (b) Uncertainty usage

Figure 8: **Comparison of subset of existing methods that adopt uncertainty estimation.**

cases after uncertainty elicitation. Their SI mitigation under S2 decreases yet remains at or above the corresponding S1 rate across $|\mathcal{V}(Q)|$. For Absent methods, SI mitigation under S2 steadily decreases, even falling below the S1 rate at $|\mathcal{V}(Q)|$=5 and reaching roughly half of it at $|\mathcal{V}(Q)|$=30, despite relatively stable detection. This suggests that methods assuming accessible model knowledge remain effective at both detecting and mitigating these hard errors robustly.

Only external knowledge mitigates CF at a meaningful rate (Fig. 7c). External knowledge methods achieve the most substantial CF and SI mitigation among the three stages, with only a limited decline as $|\mathcal{V}(Q)|$ increases. Post-hoc methods detect UNK strongly but provide almost no mitigation correction for either UNK or CF, demonstrating that detecting an abstention does not itself correct an incorrect response. Inference-time intervention achieves the highest overall detection but comparatively weak mitigation, with S2 SI mitigation declining even as SI recall increases. The distinction is therefore pronounced. Inference-time methods are effective at detecting target errors without reliably mitigating them, whereas external knowledge methods are comparatively effective at mitigation despite more moderate detection.

5.3 Uncertainty Estimation for Hallucinations

Uncertainty estimation detects UNK best but mitigates only SI (Table 3). Detection methods assign higher hallucination scores to UNK than to SI and CF. This means that the hallucination detection methods strongly align with the UNK responses, consistently stronger than the CF regardless of the uncertainty elicitation setting (i.e., both S1 and S2 labeling schemes). The gap arises as the uncertainty based methods rely on uncertainty cues, which CF responses do not carry due to their confident incorrectness.

Output-level signals are well-suited to UNK while latent representations are promising for CF (Fig. 8a). Verbal (output-level) uncertainty achieves the best detection on UNK among uncertainty representations, while showing the worst detection on CF. Latent representations achieve the best detection on CF among uncertainty representations, better than SI across all $|\mathcal{V}(\mathcal{Q})|$. The favorable patterns of latent methods hold in mitigation. Among the uncertainty representations, latent methods mitigate the most CF and UNK responses. The advantage of using verbal representation do not translate to mitigation effect, however. Instead, semantic representation shows the best mitigation correctness on **si**. Methods in the None group mitigate CF even better than latent methods, but they rely on external knowledge or decoding instead of uncertainty.

Correction methods mitigate CF and UNK slightly better than calibration methods (Fig. 8b). Mitigation methods that use uncertainty to correct the original response, achieve slightly higher mitigation effect on CF and UNK than Calibration methods, which use uncertainty primarily as a scoring signal. Mitigation methods therefore represent a promising step beyond calibration, yielding measurable gains in CF and UNK mitigation. However, these gains remain modest relative to their near ceiling detection of SI, CF, and UNK, leaving substantial room to translate successful detection into correct answers.

6 Recommendations

Based on our findings, we suggest four practical steps for hallucination research. First, **define hallucination as confident incorrectness** and explicitly distinguish it from expressed uncertainty, as conflating them risks targeting different phenomena under the same label. Second, **report CF, SI, and UNK separately** rather than aggregating all incorrect outputs, so that performance gains can be attributed to the right error type. Third, **test semantically equivalent phrasings before introducing external knowledge**: errors recoverable through rephrasing may not require retrieval augmentation, and this diagnostic can be often inexpensive relative to the interventions it can rule out. Fourth, **match uncertainty representation to the target error type**: verbal, latent, and semantic signals perform differently across each type, and latent representations are more promising than verbal ones for mitigating CF and UNK. We hope these steps make hallucination evaluations more comparable across studies and help ensure that reported improvements reflect progress on the intended problem.

7 Limitations and Conclusion

Hallucination research currently lacks consistent operational distinctions across multiple error types. The term is applied inconsistently, and behaviorally different incorrect responses are often aggregated into a single metric. We decompose incorrect model outputs into three operational categories: confabulation (CF), semantic illusion (SI), and expressed uncertainty (UNK). Empirically, we show that CF remains substantially harder for existing methods to mitigate, and to detect under uncertainty elicitation, while SI is often recoverable under inference-time intervention. Although our study covers a limited set of models, datasets, and methods, the SI–CF mitigation gap holds across both models, both labeling schemes, and all variant-set sizes. Because CF is defined by persistent incorrectness under prompt variation, one might expect prompt-level methods to fail on CF by construction. This is not the case in our setup: $\mathcal{V}(Q)$ is used only for labeling, and evaluated methods operate on held-out prompts they never see during labeling. The gap also holds for methods that do not use paraphrasing at all, including inference-time decoding and latent-state methods (Fig. 7c, 8a), so the taxonomy predicts which interventions succeed rather than restating its definition. We propose that future work define hallucination precisely, disaggregate error types, and audit alignment between definitions and evaluation methods.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data, specifically the semantically equivalent question paraphrases that make up $\mathcal{V}(Q)$, creating data visualization for our results, interpreting results, and refining and proof-reading the draft. We have reviewed all AI-assisted work. Generated paraphrases were checked for semantic equivalence and for the absence of answer hints, and we further verified that they do not increase the likelihood of the gold answer. AI-assisted interpretations were verified by the authors against the underlying experimental results and all text edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Survey responses were collected from researchers who publish on LLM hallucinations. All responses were collected before respondents were informed of our hypotheses or taxonomy, to avoid biasing them. Participation was voluntary, all respondents agreed to the use of their responses, and only aggregate results are reported. All QA data come from publicly available benchmarks (TriviaQA, SimpleQA) and contain no private information.

REFERENCES

[1] Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970, 2026. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2026.100970. URL https://www.sciencedirect.com/science/article/pii/S157401372600078X.

[2] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, Volume 8 - 2025, 2025. ISSN 2624-8212. doi: 10.3389/frai.2025.1622292. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292.

[3] N. Berlyne. Confabulation. *British Journal of Psychiatry*, 120(554):31–39, 1972. doi: 10.1192/bjp.120.554.31.

[4] Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24354–24385, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1239. URL https://aclanthology.org/2025.emnlp-main.1239/.

[5] Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models, 2026. URL https://arxiv.org/abs/2604.10697.

[6] Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Comput. Surv.*, 57(7), March 2025. ISSN 0360-0300. doi: 10.1145/3716846. URL https://doi.org/10.1145/3716846.

[7] Jennifer Chen, Aidar Myrzakhan, Yaxin Luo, Hassaan Muhammad Khan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. DRAG: Distilling RAG for SLMs from LLMs to transfer knowledge and mitigate hallucination via evidence and graph-based distillation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7240–7260, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.358. URL https://aclanthology.org/2025.acl-long.358/.

[8] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models, 2024.

[9] Yijie Hao, Haofei Yu, and Jiaxuan You. Beyond facts: Evaluating intent hallucination in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7046–7069, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.349. URL https://aclanthology.org/2025.acl-long.349/.

[10] Zheng Yi Ho, Siyuan Liang, Sen Zhang, Yibing Zhan, and Dacheng Tao. Novo: Norm voting off hallucinations with attention heads in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaOe2xBcLC.

[11] Baixiang Huang, Canyu Chen, Xiongxiao Xu, Ali Payani, and Kai Shu. Can knowledge editing really correct hallucinations? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hmDt068MoZ.

[12] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

[14] Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.187. URL https://aclanthology.org/2025.emnlp-main.187/.

[15] Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective, 2024. URL https://arxiv.org/abs/2402.06647.

[16] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 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)*, pp. 1601–1611, 2017.

[17] Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S. Yu. Loki's dance of illusions: A comprehensive survey of hallucination in large language models, 2025. URL https://arxiv.org/abs/2507.02870.

[18] Buyun Liang, Liangzu Peng, Jinqi Luo, Darshan Thaker, Kwan Ho Ryan Chan, and Rene Vidal. SECA: Semantically equivalent and coherent attacks for eliciting LLM hallucinations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8Flpo0zaaO.

[19] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

[20] Marc Marone and Benjamin Van Durme. Data portraits: Recording foundation model training data. *Advances in Neural Information Processing Systems*, 36:15121–15135, 2023.

[21] Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, and Golnoosh Farnadi. Hallucination detox: Sensitivity dropout (SenD) for large language model training. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5538–5554, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.276. URL https://aclanthology.org/2025.acl-long.276/.

[22] Morris Moscovitch. Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory. In *Varieties of memory and consciousness*, pp. 133–160. Psychology Press, 1989.

[23] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EmQSOi1X2f.

[24] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in NLP. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6528–6548, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.375. URL https://aclanthology.org/2024.emnlp-main.375/.

[25] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023. URL https://arxiv.org/abs/2309.05922.

[26] Armin Schnider. *The confabulating mind: How the brain creates reality*. Oxford University Press, 2008.

[27] Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35712–35731, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1809. URL https://aclanthology.org/2025.emnlp-main.1809/.

[28] Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. Unsupervised hallucination detection by inspecting reasoning processes. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22117–22129, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1124. URL https://aclanthology.org/2025.emnlp-main.1124/.

[29] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024. URL https://arxiv.org/abs/2401.01313.

[30] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

[31] Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics:*

ACL 2025, pp. 21381–21396, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL https://aclanthology.org/2025.findings-acl.1101/.

[32] Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. Active layer-contrastive decoding reduces hallucination in large language model generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3028–3046, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.150. URL https://aclanthology.org/2025.emnlp-main.150/.

[33] Wan Zhang and Jing Zhang. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5), 2025. ISSN 2227-7390. doi: 10.3390/math13050856. URL https://www.mdpi.com/2227-7390/13/5/856.

[34] 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. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, December 2025. doi: 10.1162/coli.a.16. URL https://aclanthology.org/2025.cl-4.9/.

Table of Contents

A Author Survey **13**

B Taxonomy Dimensions **13**

C Dataset Construction **14**

C.1 Source Datasets 14

C.2 Confidence Prompt Template 15

C.3 Behavioral Evaluation 15

A Author Survey

We surveyed 24 first authors of hallucination papers published at top-tier ML/NLP venues. Respondents could select all options that apply, so percentages do not sum to 100% in Figure 9. Each value is the fraction of the 24 respondents who selected that option. The exact values for Knowledge source and Error type are 45.8% and 41.7%. All responses were collected before respondents were informed of our hypotheses or taxonomy.

B Taxonomy Dimensions

Prompt sensitivity A method models prompt sensitivity if it explicitly evaluates how responses vary across semantically equivalent prompt variants $\mathcal{V}(Q)$. Methods testing on a single prompt do not model this variation.

Intervention stage We distinguish three points where a method can intervene: during generation (*inference*), after generation (*post-hoc*), or by augmenting the model capabilities with *external knowledge*.

Knowledge assumption A method either assumes the correct knowledge exists in the model but is misapplied, pointing to a retrieval failure (*present knowledge*), or assumes the knowledge is absent entirely and must be supplied through external sources (i.e., external knowledge), pointing to a knowledge gap (*absent knowledge*).

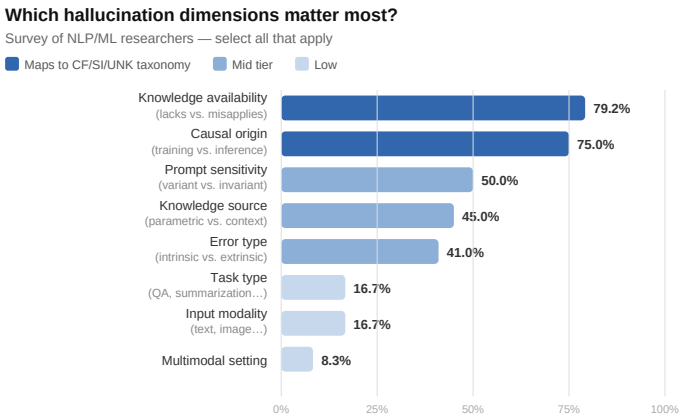


Figure 9: Hallucination survey asked to 24 first authors of hallucination papers in top-tier ML/NLP venues.

Uncertainty usage Uncertainty signals can serve two purposes, either (i) estimating how likely a response is to be correct (*calibration*), or (ii) actively modifying outputs based on that signal (*correction*)

Uncertainty representation A method may capture, express, or quantify uncertainty through expressed uncertainty in outputs (*verbal*), internal model states (*latent*), consistency across multiple outputs (*semantic*), or not engage with uncertainty at all (*none*).

Target error Each method primarily targets an error type from our taxonomy: *semantic illusion*, addressed through prompt sensitivity modeling; *confabulation*, addressed through knowledge augmentation; or *expressed uncertainty*, addressed through verbal uncertainty signals.

C Dataset Construction

This appendix provides the complete dataset construction protocol, including prompt templates, behavioral grid specification, and labeling criteria. The main paper (§4) summarizes the design; this appendix provides the detail needed for replication.

C.1 Source Datasets

Candidate questions are drawn from two publicly available factual QA benchmarks:

- **TriviaQA** [16]: open-domain trivia questions with verified answers. We use the unfiltered development split and restrict to questions with short-phrase answers to ensure unambiguous correctness evaluation.
- **SimpleQA** [30]: a benchmark specifically designed to probe factual knowledge with short, verifiable questions and a structured confidence-elicitation protocol. We adopt SimpleQA's calibration prompt (see §C.2).

We apply the following pre-filters before behavioral evaluation: (i) answers must be short and unambiguous (no date ranges, no "either X or Y" answers); (ii) questions must not require world-knowledge updated after the training cutoff of any construction model; (iii) duplicate or near-duplicate questions are removed. The remaining pool of candidates is passed to the behavioral evaluation grid described in §C.3. We ran the code in 2 $\times$ A100 40GB GPUs.

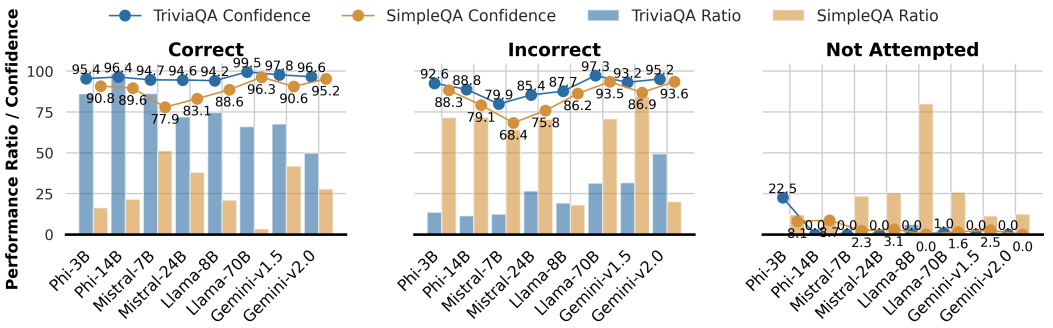


Figure 10: Model confidence scores across different response categories – correct/incorrect responses and not attempted – on the TriviaQA validation split (5k) and SimpleQA (4.3k). High confidence on incorrect responses and near-zero abstention rates confirm that abstention under the confidence prompt is a rare and therefore discriminating signal.

C.2 Confidence Prompt Template

All responses during construction and evaluation are elicited using the following prompt, adapted directly from SimpleQA's calibration protocol:

Confidence Prompt (used for all construction models)

```
Here is the question: {question}

Please provide your best guess and a confidence score between 0% to
100% in the following JSON format:

{
"answer": "Your answer here",
"confidence_score": number
}
```

This format has two deliberate properties relevant to the UNK operationalization:

1. **Abstention Signal.** The instruction "your best guess" discourages abstention. A model that still assigns `confidence_score = 0` or produces an explicit refusal is expressing stronger uncertainty than it would under open-ended prompting, because it is overriding an explicit instruction to commit. This makes UNK a conservative and meaningful signal.
2. **Abstention Detection.** The JSON structure allows precise operationalization: abstention is defined as `confidence_score = 0` *or* the answer field containing an explicit uncertainty phrase ("I don't know", "I'm not sure", "I cannot determine", "uncertain", or equivalent). No manual annotation is required.

We observe that under this prompt, models across families consistently assign high confidence to incorrect responses (Figure 10), with near-zero not-attempted rates for most model families [30]. This baseline establishes that abstention under the confidence prompt is rare by default, reinforcing its value as a discriminating signal.

C.3 Behavioral Evaluation

Each candidate questions per category NH, SI, CF, and UNK are evaluated under a dense behavioral grid designed to probe response stability across input perturbations that are semantically equivalent but surface-form variable.

Instruction variants. We evaluate each question under 4 instruction prompt variants that vary the framing of the task. Each instruction variant is crossed with each perturbation condition.

Inference-time decoding configurations. We evaluate each (question, condition, instruction) triple under multiple decoding settings: greedy decoding, temperature sampling ($T \in \{0.2, 0.7\}$), and nucleus sampling ($p = 0.80, 0.95$). This ensures that behavioral labels are not artifacts of a single decoding strategy.

Semantic paraphrase variants. We additionally evaluate each question under 30 semantically equivalent paraphrases that vary linguistic structure and lexical choices without adding hints or information.

Instruction Prompt Variants

1 `Answer the question with a short factual phrase. Output only the answer.`

2 `Return a concise answer only. No explanation.`

3 `Provide only the entity/date/name that answers the question.`

4 `Give the most likely factual answer. If uncertain, say: I don't know.` *(uncertainty-elicitation)*