Salih, Youssef Ghalib 2(2026). Fluent but Unfounded: A Review of Hallucination in Small Language Models. *SOLAV Journal*
https://solav.me/a/s2026781212

SOLAV Journal
A Scholarly Platform for Interdisciplinary Applied Futures Research
Publisher's Home Page: https://solav.me

Review Article Open Access

Fluent but Unfounded: A Review of Hallucination in Small Language Models

Salih, Youssef Ghalib[1*]

Abstract
Small language models (SLMs), compact generative models ranging from a few hundred million to roughly ten billion parameters, are increasingly chosen for on-device, private and cost-sensitive deployments. Their tendency to produce fluent but unsupported statements is nevertheless rarely treated as a distinct engineering problem. This review synthesizes research published up to November 2026 on why SLMs hallucinate, how the problem is measured, and which countermeasures exist at the model, inference and operational levels. Three arguments are made. First, the evidence points to several interacting causes, namely limited capacity for rare facts, training and evaluation incentives that reward guessing, unstable retrieval of knowledge the model does hold, and weak calibration, rather than to a single deficit. Second, reported hallucination rates depend strongly on model variant, decoding settings and task, so figures cannot be compared across studies without care. Third, no single safeguard suffices: reliability is better pursued as a layered stack that combines targeted fine-tuning, retrieval grounding, sampling-based consistency checks, abstention, and risk-proportionate human oversight. The review then maps these layers onto accountability and regulatory expectations and lists open problems, including size-controlled benchmarks and independent replication of promising small-scale methods.

Keywords
Small language models · Hallucination · Factuality · Retrieval-augmented generation · Uncertainty estimation · Abstention · AI governance · Misinformation

1. Introduction

For much of the recent history of language modelling, capability tracked scale. That assumption has loosened. Surveys of compact models document rapid gains in what systems of a few billion

* Correspondence: youssef@secria.me
1 Independent Researcher

parameters can do [1][2], and one line of argument holds that, for the repetitive and narrow calls that make up much agentic software, small models are not merely adequate but preferable on cost, latency and controllability [3]. Open-weight families such as Phi-3 [4], Gemma 2 [5], Qwen3 [6], Mistral 7B [7] and TinyLlama [8] run on a laptop, a phone or a modest server. Data can therefore stay on the device, and small organisations can adopt language technology without a cloud contract.

The same models inherit, and in places sharpen, the central weakness of the technology. A hallucination is generated text that is unfaithful to a supplied source or unsupported by verifiable fact, yet delivered in the same assured register as correct text [9][10][11]. The danger lies less in the occasional error than in the absence of an outward warning sign: a wrong answer from a three-billion-parameter model reads exactly like a right one.

Evidence specific to small models is now accumulating. A 2026 study of ten open models between 7B and 32B parameters found that one prompt frequently yields both correct and hallucinated answers across repeated samples, which the authors interpret as unstable retrieval of knowledge that the model actually holds [12]. Work on abstention in models below 10B parameters reports that small models tend to answer when they should decline, and that a cheap behavioral probe can recover part of the lost caution [13]. On the training side, a parameter-efficient method applied to the 2.7B-parameter Phi-2 model reports measurable gains in reasoning reliability [14]. These findings, however, sit in different literatures, use different definitions of “small”, and seldom speak to one another.

The stakes extend beyond benchmark scores. Machine-generated misinformation circulates widely [15]; an analysis of 91,452 posts flagged by X’s Community Notes found that AI-generated misleading posts were more likely to go viral than conventional ones [16]. Deployed systems have already produced concrete liability: a Canadian tribunal held an airline responsible for false information given by its website chatbot [17], and studies document fabricated bibliographic references [18] and high rates of legal hallucination [19][20]. Most of this evidence concerns larger models, but the lesson carries over. Harm is set by the deployment context, and the spread of SLMs extends language-model exposure into settings that often have less oversight and less logging than a managed cloud service. Accountability practices developed for other evidence-sensitive fields, such as the bias, transparency and accountability framework proposed for historical research [21], offer useful precedent.

This review makes four contributions. It (i) maps hallucination types to the mechanisms that produce them in small models, (ii) identifies pitfalls in measuring hallucination, (iii) organizes mitigation into a three-layer reliability stack, and (iv) links that stack to governance and risk-tiered deployment. The review is narrative rather than systematic. Sources were selected for relevance, venue quality and recency, with peer-reviewed work preferred and recent 2025–2026 preprints identified as such. Coverage is limited to text generation and to literature available up to November 2026.

2. Definitions, Taxonomy and Measurement

2.1 What counts as "small"

No agreed threshold exists. The contrastive fine-tuning study in [14] targets models below 3B parameters, the abstention work in [13] uses below 10B, and the geometric analysis in [12] calls 7B–32B models "small-sized". This review adopts below 10B parameters as a working range and flags evidence drawn from larger models. Many local deployments also use quantized weights, which is a further variable that comparisons often leave uncontrolled.

2.2 A working taxonomy

The general surveys distinguish factual errors, which conflict with the world, from faithfulness errors, which conflict with a source or instruction [9][10]. Two additional categories matter for small models. Confabulation denotes arbitrary wrong answers that change from one sample to the next [22], and reasoning hallucination denotes fluent multi-step solutions whose steps do not follow logically [14]. Table 1 summarizes the categories used in the rest of the review.

Type	What is wrong	Illustrative case	Usual check
Factual error	Statement contradicts verifiable fact	Wrong date or office-holder	Reference source or retrieval
Confabulation	Arbitrary answer that varies between samples	A different "winner" on each run	Sampling-based consistency [22]
Faithfulness error	Output contradicts or ignores the supplied source	Summary adds a claim absent from the document	Entailment check against the source
Reasoning error	Plausible steps that do not follow	Arithmetic chain with an invalid step [14]	Step-level verification
Fabricated reference	Non-existent citation, URL or API	Invented paper or library function [18]	Resolver, compiler or database lookup

Table 1 Hallucination types and usual detection routes

2.3 Measuring hallucination

Several benchmarks anchor the field. TruthfulQA probes whether models repeat common misconceptions [23]. HaluEval supplies generated and human-annotated samples across question answering, dialogue and summarization [24]. FActScore breaks long-form text into atomic claims and scores the share supported by a knowledge source [25]. SelfCheckGPT detects unsupported statements without external resources by comparing multiple samples [26], and HalluLens separates extrinsic from intrinsic hallucination and uses dynamically generated test sets to reduce contamination [27].

For small models, the headline number is the least informative part of a result. The geometric study in [12] illustrates why. It reports prompt-level hallucination rates between about 66% (Phi-4, 14B) and 88% (Mistral-7B) on questions about dated real-world events, but it sampled non-instruction-tuned base models at temperature 1.0 with 4-bit quantization, and it used an LLM judge whose labels matched human consensus on 87.9% of 1,000 manually checked instances. Those rates are valid for that design. They are not what a person would see from an instruction-tuned assistant at default settings. Five design choices should therefore be reported whenever hallucination rates are compared: the model variant (base or instruction-tuned), the decoding settings, the quantisation level, the judging method, and the treatment of abstentions. Paired designs that evaluate the same prompts before and after an intervention are more informative than comparisons between separately run studies.

3. Why Small Models Hallucinate

3.1 Capacity and the long tail

A model can only retain so much. Controlled experiments on synthetic knowledge estimate that transformers store roughly two bits of knowledge per parameter when training is sufficient [28], which puts a ceiling on the number of distinct facts a small model can hold. Retention is also uneven. Accuracy on a question track how often the relevant information appeared in the pretraining data, and long-tail facts are learned poorly [29]. Experiments on questions about unpopular entities show that scaling does little to close the gap, that retrieval helps considerably, and that retrieving only when the entity is rare beats always retrieving [30]. For an SLM the long tail begins earlier: facts that a large model recalls reliably may fall into the poorly retained region.

3.2 Incentives and theoretical limits

Next-token training does not by itself teach a model to say "I don't know". It has been argued that hallucinations persist because pretraining statistics produce errors on rarely seen facts and because most benchmarks grade answers as right or wrong, so that guessing is never penalized relative to abstaining [31]. Related theory shows that a calibrated model must hallucinate at a rate tied to the share of facts that appear only once in its training data [32], and a separate formal argument holds that no computable model can avoid hallucination on all inputs [33]. These results apply to all language models. Their relevance here is that they set realistic expectations: the engineering goal is reduction and containment, not elimination.

3.3 Unstable retrieval rather than missing knowledge

The assumption that every wrong answer reflects absent knowledge is incomplete. In [12], repeated sampling of the same prompt produced both correct and incorrect answers. Correct responses clustered more tightly in embedding space than hallucinated ones, and this difference supported a classifier reaching about 90% F1 from only 30–50 labelled responses. The authors read this as evidence that part of the problem lies in how knowledge is surfaced at generation time. Because the setting was base models at high temperature, the result demonstrates a mechanism more than a deployment rate, but it has a practical corollary: when a model's answers disagree with one another, the disagreement itself is informative.

3.4 Weak calibration

Under suitable formats, large models can estimate whether their own answers are correct [34], and they can be trained to verbalize calibrated confidence [35]. Confidence elicited by prompting alone, however, tends toward overconfidence [36], and the abstention work on sub-10B models reports the same difficulty together with a bias toward answering [13]. The practical consequence is that asking a small model "how sure are you?" yields a number that should not be taken at face value.

3.5 Prompt and context fragility

Pretrained models give inconsistent answers to paraphrases of the same factual query [37]. Formatting choices that a person would consider irrelevant can shift few-shot accuracy by large margins [38], and information placed in the middle of a long input is used less reliably than information at its beginning or end [39]. Small models have less capacity with which to compensate for such effects and are plausibly more exposed to them, although size-controlled comparisons remain scarce.

3.6 Side effects of the training pipeline

Fine-tuning on facts that the model did not acquire in pretraining is learned slowly and increases the tendency to hallucinate [40]. Using supervised fine-tuning to inject knowledge into a small model is therefore risky. Preference-based alignment improves instruction following and truthfulness [41], but it optimizes for what annotators prefer, which may include answers that sound complete and confident.

3.7 Fragile multi-step reasoning

Chain-of-thought prompting yields its largest gains in big models [42], and small models often produce chains that read plausibly while being logically unsound. This is the failure targeted by contrastive chain-of-thought fine-tuning, in which Phi-2 was adapted with LoRA on pairs of valid reasoning paths and explicitly labelled fallacies; the authors report 12.5% higher final-answer accuracy than standard fine-tuning on arithmetic (GSM8K) and symbolic (BBH) evaluations [14]. The result is promising but comes from a single study and one model family, so it is best read as a hypothesis awaiting independent replication.

Mechanism	What happens	Typical symptom	Evidence
Limited capacity	Fewer facts can be stored; rare facts are lost first	Errors on less popular entities	[28][29]
Training and evaluation incentives	Guessing is rewarded over abstaining	Confident wrong answers	[31][33]
Unstable retrieval	Held knowledge is not reliably surfaced	Different answers on repeated samples	[12]
Weak calibration	Stated confidence is unrelated to correctness	Overconfident errors	[13][34][36]
Prompt and context fragility	Wording or placement changes the output	Paraphrase flips the answer	[37][39]
Pipeline side effects	New-knowledge fine-tuning, preference optimization	Fabrication after adaptation	[40][41]
Fragile multi-step reasoning	Chains lose logical validity	Right-sounding but wrong derivation	[14][42]

Table 2 Mechanisms behind hallucination in small models

4. Mitigation as a Layered Reliability Stack

Because the causes in Section 3 are several, the countermeasures are several too, and they act at different points. This review groups them into three layers: a model layer that requires access to weights, an inference layer that wraps the model at run time, and an operational layer that governs how outputs are used.

4.1 Model layer

Data quality. Curating compact, high-quality training data allowed a small code model to rival much larger ones, which supports the view that for SLMs the quality of data matters more than its volume [43]. Filtering factual errors and balancing domain coverage address the capacity and long-tail problems of Section 3.1 at their source.

Parameter-efficient adaptation. LoRA and QLoRA make domain adaptation feasible on a single GPU [44][45], which is the realistic compute budget for most SLM stakeholders. Adaptation should be used to teach behavior, format and domain style rather than to inject facts the model has not rehearsed, because new-knowledge fine-tuning raises hallucination risk [40].

Preference and reasoning objectives. Reinforcement learning from human feedback [41] and the simpler direct preference optimization [46] can include honesty and abstention preferences. For multi-step tasks, contrastive chain-of-thought fine-tuning pairs valid reasoning with labelled fallacies [14]; the reported 12.5% gain over standard fine-tuning is encouraging but still awaits independent confirmation on other model families and on free-form factual tasks.

Abstention-aware training and decoding. Uncertainty-based abstention has been shown to improve correctness and reduce hallucination [47], and a survey organizes the many ways a model can be taught or prompted to decline [48]. Broader catalogues of mitigation methods are available [49].

4.2 Inference layer

Retrieval grounding. Retrieval-augmented generation supplies evidence at generation time [50][51]. It suits SLMs particularly well, since it moves facts out of the parameters, where capacity is scarce, and into an index that can be updated, and selective retrieval for rare entities is both effective and economical [30]. Retrieval is not a cure: poor passages produce faithfulness errors,

and commercial retrieval-based legal research tools still gave incorrect or misgrounded answers in a noticeable share of queries [20].

Consistency and uncertainty from sampling. Self-consistency aggregates several reasoning paths and keeps the majority answer [52]. SelfCheckGPT flags statements that disagree across samples [26], and semantic entropy measures uncertainty over meanings rather than wordings so as to detect confabulation [53][22]. The geometric signature in [12] extends this idea, because tightly clustered responses indicate reliability. The price is several generations per query, which matters for latency and battery on edge devices.

Cheap behavioral probes. Second Guess asks a multiple-choice question twice, the second time with an extra "I don't know" option, and treats a changed answer as a signal to abstain. Across four open models of 2B–8B parameters and four benchmarks it lowered composite risk by about 10.8 percentage points on average using only two queries per question [13]. The gains came mostly from detecting answer switching rather than from the models actually choosing the abstention option, and the evaluation was multiple-choice, so transfer to free-form generation is an open question.

Verification pipelines. Chain-of-verification drafts an answer, plans verification questions, answers them independently of the draft, and then revises [54]. The pattern can be implemented around an SLM at modest cost, using the same model or a second one as verifier.

4.3 Operational layer

The last layer concerns what people and organizations do with the output. Useful practices include requiring answers to quote the supplied context; asking for explicit uncertainty while treating the stated confidence as a weak signal (Section 3.4); decomposing complex requests into smaller, checkable questions; validating code with compilers and tests; resolving cited references against bibliographic databases, since fabricated citations are a documented failure [18]; and routing consequential outputs to human review. Agreement between several small models is a useful but imperfect screen, because models trained on overlapping data may share the same errors. Logging prompts, outputs and verification outcomes turns individual catches into evidence for improving the system.

Layer	Technique	Needs model weights	Main benefit	Main limit	Refs

Model	Curated data, LoRA/QLoRA, preference optimization	Yes	Shifts the behavior at its source	Cannot add unrehearsed facts safely	[40][43][46]
Model	Contrastive chain-of-thought fine-tuning	Yes	Targets unsound reasoning	Single-study evidence	[14]
Model	Abstention-aware training	Yes	Reduces confident errors	Risk of over-refusal	[47][48]
Inference	Retrieval grounding	No	Moves facts out of parameters	Retrieval and grounding errors	[20][30][50][51]
Inference	Sampling consistency, semantic entropy	No	Detects confabulation	Several generations per query	[12][26][52][53]
Inference	Two-pass abstention probe	No	Very low cost	Tested on multiple-choice tasks	[13]
Inference	Chain-of-verification	No	Checks claims independently	Verifier may share the errors	[54]
Operational	Source quoting, tool validation, human review	No	Independent of the model	Time and expertise	[18]

Table 3 The three layers compared

5. Governance, Accountability and Societal Risk

5.1 Hallucination as an information-ecosystem risk

Hallucinated text does not stay where it is generated. Machine-written misinformation spreads through the same channels as other false content [15], and the empirical evidence from X suggests that AI-generated misleading posts attract disproportionate engagement even though they are judged slightly less believable [16]. A plausible concern, not yet well measured, is that locally run SLMs weaken platform-level safeguards, because outputs are produced and shared outside any moderated or logged service.

5.2 Who is accountable

The airline decision in [17] is instructive. The tribunal rejected the argument that a chatbot is a separate entity answerable for its own statements and held the deployer responsible for what the chatbot told a customer. Responsibility therefore rests with those who deploy a system, not with the model. Sector studies reinforce the point: fabricated references in scholarly writing [18] and

high legal hallucination rates [19][20] show that harms arise where users cannot easily check claims. The accountability framework developed for historical research, built on bias mitigation, transparency and accountability [21], illustrates how an evidence-sensitive discipline can formalize duties that transfer to SLM deployments: document provenance, disclose limitations, and assign a named owner for verification.

Regulatory instruments point in the same direction. The NIST generative AI profile names confabulation as a distinct risk and proposes actions to manage it [55]. The EU Artificial Intelligence Act scales obligations with risk and, for high-risk systems, requires measures on accuracy, robustness and human oversight [56]. Documented hallucination controls such as those in Section 4 are natural evidence of compliance with such expectations.

5.3 Bias and amplification

Small models trained on narrower corpora can reproduce and amplify the social biases present in their data, and the scale and cost of training large corpora carry their own ethical questions [57]. Unchecked hallucination compounds the problem, because a fabricated claim about a person or group is harder to challenge when it arrives in fluent prose.

5.4 Risk-proportionate deployment

Safeguards cost time, so they should scale with consequence. Table 4 offers a simple tiering that combines the layers of Section 4.

Tier	Example uses	Minimum safeguards
Low	Brainstorming, style editing, drafts without factual claims	Disclosure that output may be wrong; user judgement
Medium	Internal summarization, customer FAQ, coding assistance	Retrieval grounding; source quoting; automated tests; logging; sampled human audit
High	Health, legal, financial, educational assessment, public information	All of the above, plus abstention thresholds, independent verification of every material claim, named accountable owner, regular re-evaluation

Table 4 Risk tiers and minimum safeguards

6. Open Problems and Research Agenda

1. Size-controlled benchmarks. Benchmarks that vary parameter count, variant, quantization and decoding independently are needed, so that the effect of "smallness" can be separated from the effect of the setup (Section 2.3).

2. Independent replication. Promising small-scale methods, including contrastive chain-of-thought fine-tuning [14] and two-pass abstention probing [13], should be replicated across model families and extended from benchmark tasks to free-form factual generation.

3. Affordable uncertainty on the edge. Sampling-based detectors [12][26][53] multiply inference cost. Methods that approximate their signal within a single pass or a small number of passes would make them practical on phones and embedded devices.

4. Retrieval for constrained devices. Compact indexes, and policies for deciding when to retrieve [30], are needed so that grounding does not erase the efficiency and privacy advantages that motivate SLMs.

5. Composition of layers. Most studies test one technique in isolation. Ablations that combine fine-tuning, retrieval, consistency checks and abstention would show which combinations add value and which duplicate each other.

6. Human factors. Evidence on how people actually verify small-model output, and on whether tiered controls such as Table 4 are followed in practice, is thin.

7. Auditability of local models. Governance regimes assume logs and responsible parties [55][56]; lightweight, privacy-preserving audit trails for locally run models remain to be designed.

7. Conclusion

Small language models make capable language technology cheap, private and widely available, and hallucination is the main obstacle to using them responsibly. The literature reviewed here suggests that the problem is best understood as the product of several mechanisms: limited capacity for rare facts, incentives that reward guessing, unstable retrieval of what the model does know, poor calibration, and fragile reasoning. Reported rates are highly sensitive to experimental design, so numbers should be read with their setup in view. Mitigation works best as a stack. Targeted fine-tuning shapes behavior, retrieval and consistency checks contain errors at run time, and operational controls and human oversight, scaled to the consequence of a mistake, decide what finally reaches a reader. Several of the most promising techniques rest on single studies, and independent replication is the most useful next step for the field.

8. Declaration

8.1 Availability of data and material

Not applicable.

8.2 Funding

Not applicable.

8.3 Acknowledgements

The author used AI tools to assist with language editing, grammar, and readability to improve clarity and precision.

References

1. Z. Lu et al., “Small language models: Survey, measurements, and insights,” arXiv:2409.15790, 2024. https://doi.org/10.48550/arXiv.2409.15790
2. F. Wang et al., “A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness,” ACM Trans. Intell. Syst. Technol., vol. 16, no. 6, pp. 1–87, 2025.
3. P. Belcak et al., “Small language models are the future of agentic AI,” arXiv:2506.02153, 2025. https://doi.org/10.48550/arXiv.2506.02153
4. M. Abdin et al., “Phi-3 technical report: A highly capable language model locally on your phone,” arXiv:2404.14219, 2024. https://doi.org/10.48550/arXiv.2404.14219
5. Gemma Team, “Gemma 2: Improving open language models at a practical size,” arXiv:2408.00118, 2024. https://doi.org/10.48550/arXiv.2408.00118
6. A. Yang et al., “Qwen3 technical report,” arXiv:2505.09388, 2025. https://doi.org/10.48550/arXiv.2505.09388
7. A. Q. Jiang et al., “Mistral 7B,” arXiv:2310.06825, 2023. https://doi.org/10.48550/arXiv.2310.06825
8. P. Zhang, G. Zeng, T. Wang, and W. Lu, “TinyLlama: An open-source small language model,” arXiv:2401.02385, 2024. https://doi.org/10.48550/arXiv.2401.02385
9. Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Comput. Surv., vol. 55, no. 12, Art. no. 248, pp. 1–38, 2023. https://doi.org/10.1145/3571730

10. L. Huang et al., “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” ACM Trans. Inf. Syst., vol. 43, no. 2, Art. no. 42, 2025. https://doi.org/10.1145/3703155
11. Y. Zhang et al., “Siren’s song in the AI ocean: A survey on hallucination in large language models,” arXiv:2309.01219, 2023. https://doi.org/10.48550/arXiv.2309.01219
12. E. Ricco, E. Onofri, L. Cima, S. Cresci, and R. Di Pietro, “A geometric analysis of small-sized language model hallucinations,” arXiv:2602.14778, 2026; presented at Int. Conf. Mach. Learn. (ICML), 2026. https://doi.org/10.48550/arXiv.2602.14778
13. A. V. Aravindan and M. Kejriwal, “Second Guess: Detecting uncertainty through abstention and answer stability in small language models,” arXiv:2605.25394, 2026. https://doi.org/10.48550/arXiv.2605.25394
14. M. A. Baker and F. Al-Qrize, “Mitigating hallucination in small language models via contrastive chain-of-thought fine-tuning,” SOLAV J., vol. 1, pp. 33–53, Jul. 2025. https://doi.org/10.5281/zenodo.18538736
15. M. A. Baker, “The spread of AI-generated misinformation,” Authorea, preprint, Apr. 7, 2025. https://doi.org/10.22541/au.174405164.43017701/v1
16. C. Drolsbach and N. Pröllochs, “Characterizing AI-generated misinformation on social media,” arXiv:2505.10266, 2025. https://doi.org/10.48550/arXiv.2505.10266
17. Moffatt v. Air Canada, 2024 BCCRT 149 (Civil Resolution Tribunal of British Columbia, Feb. 14, 2024). https://canlii.ca/t/k2b02
18. W. H. Walters and E. I. Wilder, “Fabrication and errors in the bibliographic citations generated by ChatGPT,” Sci. Rep., vol. 13, Art. no. 14045, 2023. https://doi.org/10.1038/s41598-023-41032-5
19. M. Dahl, V. Magesh, M. Suzgun, and D. E. Ho, “Large legal fictions: Profiling legal hallucinations in large language models,” J. Legal Anal., vol. 16, no. 1, pp. 64–93, 2024.
20. V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, and D. E. Ho, “Hallucination-free? Assessing the reliability of leading AI legal research tools,” J. Empir. Leg. Stud., vol. 22, no. 2, pp. 216–242, 2025.
21. M. A. Baker, “AI ethics in historical research: A framework for bias mitigation, transparency, and accountability,” SSRN, Oct. 5, 2025. https://doi.org/10.2139/ssrn.5565499
22. S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” Nature, vol. 630, no. 8017, pp. 625–630, 2024. https://doi.org/10.1038/s41586-024-07421-0

23. S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2022, pp. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229
24. J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A large-scale hallucination evaluation benchmark for large language models,” in Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP), 2023, pp. 6449–6464. https://doi.org/10.18653/v1/2023.emnlp-main.397
25. S. Min et al., “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in Proc. EMNLP, 2023, pp. 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741
26. P. Manakul, A. Liusie, and M. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” in Proc. EMNLP, 2023, pp. 9004–9017. https://doi.org/10.18653/v1/2023.emnlp-main.557
27. Y. Bang et al., “HalluLens: LLM hallucination benchmark,” in Proc. 63rd Annu. Meeting ACL, 2025, pp. 24128–24156.
28. Z. Allen-Zhu and Y. Li, “Physics of language models: Part 3.3, knowledge capacity scaling laws,” arXiv:2404.05405, 2024. https://doi.org/10.48550/arXiv.2404.05405
29. N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large language models struggle to learn long-tail knowledge,” in Proc. 40th Int. Conf. Mach. Learn. (ICML), 2023, pp. 15696–15707. https://proceedings.mlr.press/v202/kandpal23a.html
30. A. Mallen et al., “When not to trust language models: Investigating effectiveness of parametric and non-parametric memories,” in Proc. 61st Annu. Meeting ACL, 2023, pp. 9802–9822. https://doi.org/10.18653/v1/2023.acl-long.546
31. A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, “Why language models hallucinate,” arXiv:2509.04664, 2025. https://doi.org/10.48550/arXiv.2509.04664
32. A. T. Kalai and S. S. Vempala, “Calibrated language models must hallucinate,” in Proc. 56th Annu. ACM Symp. Theory Comput. (STOC), 2024, pp. 160–171.
33. Z. Xu, S. Jain, and M. Kankanhalli, “Hallucination is inevitable: An innate limitation of large language models,” arXiv:2401.11817, 2024. https://doi.org/10.48550/arXiv.2401.11817
34. S. Kadavath et al., “Language models (mostly) know what they know,” arXiv:2207.05221, 2022. https://doi.org/10.48550/arXiv.2207.05221
35. S. Lin, J. Hilton, and O. Evans, “Teaching models to express their uncertainty in words,” Trans. Mach. Learn. Res., 2022. https://openreview.net/forum?id=8s8K2UZGTZ
36. M. Xiong et al., “Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2024. https://openreview.net/forum?id=gjeQKFxFpZ

37. Y. Elazar et al., “Measuring and improving consistency in pretrained language models,” Trans. Assoc. Comput. Linguistics, vol. 9, pp. 1012–1031, 2021. https://doi.org/10.1162/tacl_a_00410
38. M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr, “Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting,” in Proc. ICLR, 2024. https://openreview.net/forum?id=RIu5lyNXjT
39. N. F. Liu et al., “Lost in the middle: How language models use long contexts,” Trans. Assoc. Comput. Linguistics, vol. 12, pp. 157–173, 2024. https://doi.org/10.1162/tacl_a_00638
40. Z. Gekhman et al., “Does fine-tuning LLMs on new knowledge encourage hallucinations?” in Proc. EMNLP, 2024. https://doi.org/10.48550/arXiv.2405.05904
41. L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, 2022, pp. 27730–27744. https://doi.org/10.48550/arXiv.2203.02155
42. J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Proc. NeurIPS, vol. 35, 2022, pp. 24824–24837. https://doi.org/10.48550/arXiv.2201.11903
43. S. Gunasekar et al., “Textbooks are all you need,” arXiv:2306.11644, 2023. https://doi.org/10.48550/arXiv.2306.11644
44. E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” in Proc. ICLR, 2022. https://openreview.net/forum?id=nZeVKeeFYf9
45. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient finetuning of quantized LLMs,” in Proc. NeurIPS, vol. 36, 2023. https://doi.org/10.48550/arXiv.2305.14314
46. R. Rafailov et al., “Direct preference optimization: Your language model is secretly a reward model,” in Proc. NeurIPS, vol. 36, 2023. https://doi.org/10.48550/arXiv.2305.18290
47. C. Tomani, K. Chaudhuri, I. Evtimov, D. Cremers, and M. Ibrahim, “Uncertainty-based abstention in LLMs improves safety and reduces hallucinations,” arXiv:2404.10960, 2024. https://doi.org/10.48550/arXiv.2404.10960
48. B. Wen et al., “Know your limits: A survey of abstention in large language models,” Trans. Assoc. Comput. Linguistics, vol. 13, pp. 529–556, 2025.
49. S. M. T. I. Tonmoy et al., “A comprehensive survey of hallucination mitigation techniques in large language models,” arXiv:2401.01313, 2024. https://doi.org/10.48550/arXiv.2401.01313
50. P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Proc. NeurIPS, vol. 33, 2020, pp. 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

51. Y. Gao et al., “Retrieval-augmented generation for large language models: A survey,” arXiv:2312.10997, 2023. https://doi.org/10.48550/arXiv.2312.10997
52. X. Wang et al., “Self-consistency improves chain of thought reasoning in language models,” in Proc. ICLR, 2023. https://doi.org/10.48550/arXiv.2203.11171
53. L. Kuhn, Y. Gal, and S. Farquhar, “Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation,” in Proc. ICLR, 2023. https://openreview.net/forum?id=VD-AYtP0dve
54. S. Dhuliawala et al., “Chain-of-verification reduces hallucination in large language models,” arXiv:2309.11495, 2023. https://doi.org/10.48550/arXiv.2309.11495
55. National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1, Jul. 2024. https://doi.org/10.6028/NIST.AI.600-1
56. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Off. J. Eur. Union, Jul. 12, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
57. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?,” in Proc. ACM Conf. Fairness, Accountability, Transparency (FAccT), 2021, pp. 610–623. https://doi.org/10.1145/3442188.3445922