# Why One Question Can Yield Many Answers: Structural Pathways to Hallucination in Transformer-Based Generative AI and a Literature-as-Data Synthesis of Expert–Non-Expert Responses

**Anonymous Author(s)**
Affiliation
Address
email

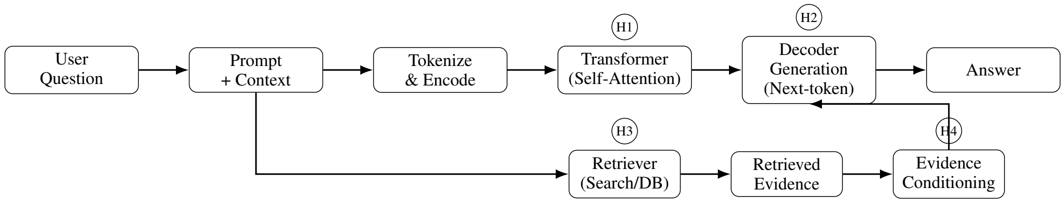
## Abstract

Hallucinations—confidently stated but false outputs—remain a major reliability barrier for large language models (LLMs). A key symptom is that even when a question has a single correct answer, an LLM may produce multiple incompatible answers across runs or conditions, and may commit to an incorrect one without calibrated uncertainty. We explain how Transformer attention and decoder-style next-token generation yield a distribution over continuations rather than guaranteed fact retrieval [10, 4], and how retrieval-augmented generation (RAG) can reduce but not eliminate hallucinations by conditioning on external evidence [5, 3]. We then treat the research literature as a corpus, extracting definitions, figures/tables, and empirical findings to code results along four axes: (1) structural mechanisms and "inevitability" arguments, (2) expert–non-expert usage patterns (proxied by education/occupation and work vs. non-work context), (3) mitigation and Human-in-the-Loop (HITL) practices, and (4) attitudes toward hallucination. Our synthesis highlights incentive misalignment that rewards guessing over calibrated abstention [4], persistent deficits in truthfulness under scaling [6], and human over-trust when explanations are provided by default [8]. We conclude by discussing specialized small language models (SLMs) and verification-centered workflows as complementary paths [1, 7].

## 1 Introduction

Generative AI systems are increasingly used in decision support, writing, and knowledge work. Yet hallucinations remain difficult to eliminate: models can produce plausible but false statements and deliver them confidently [4]. This paper focuses on a practical and structurally grounded framing: *"one question, many answers"*—a state where the same query admits multiple incompatible model outputs, and the system may select an incorrect one without calibrated uncertainty. We aim to (i) describe where this state arises in the LLM pipeline (attention/decoder/sampling; retrieval; evaluation incentives), and (ii) synthesize prior empirical results on how users respond to hallucination, with an emphasis on expert–non-expert differences and HITL checkpoints [2, 8].

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.




**Hallucination risk points:** H1: representation $\neq$ fact retrieval [10]. H2: sampling/decoding commits to a plausible continuation [4]. H3: retrieval failure or irrelevant context [5]. H4: unfaithful use of evidence / citation mismatch [3].

Figure 1: LLM pipeline with typical hallucination risk points. The diagram is an explanatory schematic grounded in core mechanisms discussed in prior work [10, 4, 5, 3].

# 2 Background: Where "One Question, Many Answers" Comes From

## 2.1 Transformer attention and representation (not retrieval)

Transformers compute context-dependent representations through self-attention, enabling each token to attend to all others [10]. This mechanism supports flexible contextual reasoning, but it does not guarantee deterministic retrieval of a single ground-truth fact.

## 2.2 Decoder-style next-token prediction and sampling

Most LLM deployments generate text autoregressively. At each step, the model outputs a probability distribution over the next token. Different decoding choices (temperature/top-$p$) or slight contextual changes can shift outputs, creating multiple plausible continuations for the same question. Even with conservative decoding, variability can arise from hidden context (system prompts), retrieval variation, or small prompt differences; with temperature/top-$p$ decoding, the distributional nature of generation becomes explicit, making multiple incompatible answers more likely. OpenAI argues that hallucinations persist partly because standard training/evaluation reward guessing over acknowledging uncertainty [4].

## 2.3 RAG: grounded generation, not a truth guarantee

RAG conditions generation on retrieved external documents, improving factual grounding in knowledge-intensive tasks [5]. However, RAG can still fail through retrieval errors, evidence misinterpretation, or unfaithful generation. RAGAS highlights the need to evaluate retrieval quality and faithfulness separately [3].

# 3 Problem Definition

We adopt an operational definition aligned with OpenAI: hallucination is *confident commitment to a false claim* [4].

**Scope and exclusions.** To keep the study scientifically tractable, we focus on *verifiable factual claims* and *provenance-bearing outputs* (citations, quoted evidence, or source-dependent specifics). We explicitly exclude (i) harmless paraphrasing or stylistic variation, (ii) open-ended normative disagreements, and (iii) creative writing where multiple answers are valid by design. Our target failure mode is therefore *confident commitment to a false, checkable claim*, including fabricated provenance [4]. We code a response as hallucination if it exhibits at least one of: (i) false factual claim with strong confidence language, (ii) correction resistance after challenge, (iii) fabricated support (invented citations or unverifiable specifics). This definition targets the failure mode "the model confirms an incorrect answer."

# 4 Method: Literature as Data (Collection and Coding)

## 4.1 Corpus construction

Rather than collecting unverified anecdotes, we treat the research literature itself as a dataset. We compile a corpus of $N = 10$ representative sources spanning: (i) core Transformer/LLM mechanisms, (ii) hallucination/truthfulness evaluation, (iii) RAG and faithfulness assessment, and (iv) human behavior and usage patterns. Our goal is coverage across mechanisms, measurement, and human factors rather than exhaustive review.

**Inclusion criteria.** A source is included if it (1) explicitly discusses hallucination, truthfulness, faithfulness, calibration, RAG, or HITL, and (2) provides at least one of: a figure/table, a benchmark protocol, or an empirical dataset/analysis. This yields a mixed corpus of benchmarks [6], technical analyses [4, 11], RAG methods/evaluation [5, 3], user studies and observational usage evidence [8, 2], and surveys/technical reports that summarize mitigation and small-model deployment directions [9, 1, 7].

## 4.2 What we extract as "data"

From each source, we extract: (a) the operational definition(s) of hallucination/truthfulness/faithfulness if provided, (b) any figures/tables and the associated reported metrics or qualitative claims, (c) described mitigation mechanisms and evaluation dimensions, (d) evidence type (theory/benchmark/user study/observational analysis/technical report/survey), and (e) explicit or implied HITL checkpoints.

## 4.3 Coding scheme (four axes)

Each source is coded along four axes (multi-label allowed):

- **A1 Structural mechanisms / inevitability:** next-token prediction limits, representation vs. retrieval, and arguments that hallucination is structurally difficult to eliminate [4, 11].
- **A2 Expert–non-expert response patterns:** proxied by work vs. non-work context and occupational/educational distributions; interpreted as *accountability context* rather than ground-truth expertise [2].
- **A3 Mitigation and HITL practices:** RAG, verification, calibration, faithfulness evaluation, and workflow checkpoints [5, 3, 9].
- **A4 Attitudes toward hallucination:** stances framing hallucination as inevitable vs. manageable via abstention, evaluation redesign, and user-side verification norms [4, 8].

## 4.4 Reliability and reproducibility

To reduce subjective drift, we used a two-pass coding procedure: a first pass to extract claims and artifacts, and a second pass (after a time gap) to re-check axis assignment and ensure internal consistency. Where feasible, future iterations can add a second coder on a held-out subset to report inter-rater agreement (e.g., Cohen's $\kappa$).

- **A1 Structural mechanisms / inevitability:** ...
- **A2 Expert–non-expert use patterns:** ...
- **A3 Mitigation and HITL practices:** ...
- **A4 Attitudes toward hallucination:** ...

# 5 Results: Synthesized Findings

**Quantitative summary of coded coverage.** Across the $N = 10$-source corpus, structural-mechanism discussions (A1) appear in 5/10 sources, expert/non-expert usage proxies (A2) in 2/10, mitigation/HITL mechanisms (A3) in 6/10, and attitude/stance discussions (A4) in 6/10 (Table 2).

| Axis | What we extract as "data" | Example sources |
|---|---|---|
| A1 | Claims about structural causes; incentive arguments; mechanisms tied to next-token prediction | [4, 10] |
| A2 | Usage distributions (work/non-work), occupation/education proxies, observed behavior | [2] |
| A3 | Mitigation taxonomy (RAG, verification, calibration); evaluation dimensions | [5, 3, 9] |
| A4 | Stances on inevitability vs. manageability; recommended norms | [4, 8] |

Table 1: Coding scheme (literature-as-data). This table is intended to make the "data collection/analysis" sections explicit and reproducible.

| Source | A1 | A2 | A3 | A4 | Evidence |
|---|---|---|---|---|---|
| Vaswani et al. (2017) [10] | ✓ | | | | Mechanism |
| OpenAI (2025) [4] | ✓ | | ✓ | ✓ | Analysis |
| Xu et al. (2024) [11] | ✓ | | | ✓ | Theory/Analysis |
| Lewis et al. (2020) [5] | ✓ | | ✓ | | Method |
| Es et al. (2023) [3] | | | ✓ | | Evaluation |
| Lin et al. (2021) [6] | ✓ | | | ✓ | Benchmark |
| NMI (2024) [8] | | ✓ | ✓ | ✓ | User study |
| NBER (2025) [2] | | ✓ | | | Observational |
| Tonmoy et al. (2024) [9] | | | ✓ | ✓ | Survey |
| Phi-3 (2024); SLM survey (2025) [1, 7] | | | ✓ | ✓ | Tech report/Survey |

Table 2: Literature-as-data coding results (multi-label). "Expert–non-expert" is treated as accountability context proxied by usage setting, not a ground-truth expertise label [2].

This distribution matches a common pattern in the field: mechanisms and mitigations are widely documented, while systematic evidence about user groups and accountability contexts is comparatively thinner.

### 5.1 Structural pathway: incentives + uncertainty suppression

OpenAI argues that benchmarks focused on accuracy can reward guessing over abstaining, making confident errors persist [4]. This explains why "one question, many answers" becomes risky specifically when the system commits to one plausible completion without calibrated uncertainty.

### 5.2 Truthfulness does not monotonically improve with scale

TruthfulQA shows that larger models can be *less truthful* on questions designed to elicit imitative falsehoods [6]. This supports the idea that fluency can increase persuasive wrong answers.

### 5.3 Human over-trust under default explanations

A Nature Machine Intelligence study identifies a calibration gap: people overestimate the accuracy of LLM responses when default explanations are provided [8]. This is a key risk amplifier for non-experts and time-pressured experts alike.

**HITL checkpoints (deployment-oriented).** For tasks where a single correct fact is required or downstream risk is high, prioritize *verification over fluency*:

1. **Single-fact queries:** require citation/evidence display or abstain ("I don't know") [4].
2. **High-stakes domains:** mandate cross-checking with $\geq$2 independent sources (human sign-off).
3. **Strong certainty language:** trigger a "verification mode" (retrieve, quote, and align claim-to-evidence) [8].
4. **RAG enabled:** evaluate retrieval relevance and faithfulness separately; check citation-to-claim alignment [3].
5. **Decision use:** record provenance (sources, prompts, retrieval snapshot) and log corrections for future audits.

Figure 2: A practical HITL checklist derived from the synthesized mechanisms and human-factor findings [4, 8, 3].

### 5.4 Expert–non-expert proxy: work usage and accountability context

NBER reports that work-related usage is more common among educated users in highly paid professional occupations [2]. While this is an imperfect proxy for "expertise," it suggests that verification norms and downstream accountability can vary by context.

## 6 Failure-Case Analysis: Where "One Question" Turns Into Hallucination

### 6.1 Pattern 1: Single-fact queries under sparse knowledge

For arbitrary low-frequency facts (dates, titles, identifiers), next-token prediction may yield multiple plausible answers; incentive structures can encourage commitment [4]. **HITL implication:** require abstention or verification checks when the query demands a single correct fact.

### 6.2 Pattern 2: Explanations amplify confidence more than correctness

Default explanations can increase perceived reliability without improving accuracy [8]. **HITL implication:** require evidence display, uncertainty cues, and user-side cross-checking for high-stakes tasks.

### 6.3 Pattern 3: RAG provenance illusion

RAG reduces hallucination risk but does not guarantee faithfulness [5, 3]. **HITL implication:** verify citation-to-claim alignment; evaluate retrieval and faithfulness separately.

## 7 Discussion: SLM Specialization and a Verification-Centered Future

### 7.1 Why specialized SLMs are promising

Small Language Models (SLMs) are attractive for efficiency, on-device deployment, and domain specialization [7]. Phi-3 demonstrates strong performance for a relatively small model, motivating a "specialized SLM + verification workflow" design space [1]. In settings where governance and predictability matter, smaller specialized models with explicit abstention policies can be easier to control than general-purpose LLMs [4].

### 7.2 Work will reorganize around verification and responsibility

Large-scale usage evidence suggests growth in both work and non-work use, with work usage concentrated among educated professionals [2]. At the same time, miscalibrated trust implies that safer deployment increasingly depends on people who can audit, cross-check, and iteratively correct model outputs [8]. Thus, rather than a simple "jobs disappear" story, an alternative trajectory is increased demand for roles centered on continuous learning, error correction, and accountability.

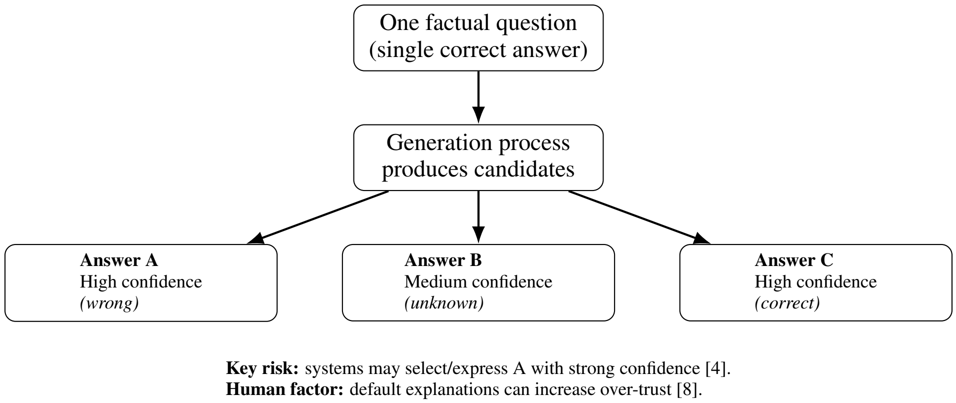



Figure 3: "One question, many answers" as a precondition for hallucination. When uncertainty is not communicated and incentives favor guessing, the system may confidently commit to an incorrect candidate [4].

# 8 Limitations

This study synthesizes existing findings rather than running new controlled experiments. The expert–non-expert distinction is approximated using proxies (occupation/education and work-context usage) [2]. Finally, explanatory figures are schematic: they illustrate mechanisms supported by cited sources but do not reproduce any single paper's original figure. We also note potential publication bias (successful mitigations are more likely to be reported than failures), and heterogeneity in reported metrics that prevents uniform effect-size aggregation across studies. A natural next step is a small controlled experiment that operationalizes "one question, many answers" under fixed decoding and retrieval conditions, enabling direct measurement of variance and calibration.

# 9 Conclusion

Transformer-based LLMs generate from distributions rather than guaranteed truth retrieval. When evaluation incentives reward guessing and uncertainty is not communicated, "one question, many answers" can become confident commitment to a false claim [4]. Literature-as-data synthesis suggests that effective mitigation is multi-stage: grounding (RAG), faithfulness evaluation, calibrated uncertainty, and explicit HITL checkpoints [5, 3, 8]. Finally, SLM specialization and verification-centered workflows offer a promising complementary direction [1, 7].


# Acknowledgments

(Optional; can exceed 9 pages along with references per template instructions.)


# References


[1] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. *NBER Working Paper 34255*, 2025. DOI: 10.3386/w34255.

[3] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.

[4] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang, et al. Why language models hallucinate, 2025.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Taras Kucheryavyy, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020.

[6] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2021.

[7] Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, Junda Wu, Ashish Singh, Yu Wang, Jiuxiang Gu, Nesreen K. Ahmed, Nedim Lipka, Ruiyi Zhang, Xiang Chen, Tong Yu, Sungchul Kim, Hanieh Deilamsalehy, Namyong Park, Michael Rimer, Zhehao Zhang, Huanrui Yang, Puneet Mathur, Gang Wu, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 807–821, 2025.

[8] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W. Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231, 2025. Published: 21 January 2025.

[9] 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.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] Rui Xu, Vinija Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024.

# AI Co-Scientist Challenge Korea paper checklist

### Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
Answer: [Yes] Justification: The abstract and introduction state that the paper provides a structural explanation of hallucination pathways and a literature-as-data synthesis; these are supported by Sections 2–7 (mechanisms, coding results, and HITL checklist).

### Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?
Answer: [Yes] Justification: Section 9 (Limitations) discusses the absence of new controlled experiments, proxy-based expert/non-expert interpretation, publication bias, and metric heterogeneity, and outlines concrete future work.

### Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
Answer: [N/A] Justification: The paper does not present new theorems or formal proofs; it synthesizes existing theoretical and empirical findings from prior work.

### Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
Answer: [N/A] Justification: The paper's main contribution is a literature-as-data coding synthesis rather than new computational experiments; reproducibility is addressed through the documented corpus, coding axes, and two-pass coding procedure in Section 4.

**Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
Answer: [No] Justification: We do not release a dedicated codebase in this submission; however, the "data" consists of publicly available papers cited in the References, and the extraction/coding procedure is described in Section 4.

**Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
Answer: [N/A] Justification: No new model training or benchmark experiments are conducted; the study is based on literature coding and synthesis.

**Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
Answer: [N/A] Justification: The paper does not report new experimental measurements; it summarizes prior results and provides a qualitative/structured synthesis.

**Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: [N/A] Justification: The paper does not include computational experiments requiring specified compute resources.

**Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://nips.cc/public/EthicsGuidelines?
Answer: [Yes] Justification: The work uses only publicly available literature and does not collect human-subject data or personal information; ethical considerations and limitations are discussed (Sections 7–9).

**Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes] Justification: Section 8 discusses potential benefits (verification-centered workflows, SLM specialization) and risks (over-trust, deployment failure in high-stakes settings), and connects them to HITL safeguards (Section 7).

**Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [N/A] Justification: We do not release new models or scraped datasets; the paper provides workflow-level safeguards (HITL checklist) for responsible use rather than a release protocol.

**Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [No] Justification: We cite all original papers and URLs for the referenced assets, but we

do not systematically list the license terms for each artifact; this can be added in a supplemental appendix if required.

**New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [N/A] Justification: The paper does not introduce or release new datasets, code, or models.

**Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [N/A] Justification: The paper does not conduct crowdsourcing or recruit human participants.

**Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review ...) were obtained?
Answer: [N/A] Justification: The paper does not involve human-subject research or collection of personal data, so IRB review is not applicable.