Evaluating the Impact of Prompt Engineering on Factual Accuracy and Hallucination in Large Language Models

Diya Jain., Pallak Anand., Dr. Deepti Sharma

Masters of Computer Applications, Jagan Institute of Management Studies, Rohini, Delhi – 110085, India

DOI: https://doi.org/10.51244/IJRSI.2026.1304000068

Received: 06 April 2026; Accepted: 12 April 2026; Published: 30 April 2026

ABSTRACT

The propensity of large language models (LLMs) to generate factually unsupported yet linguistically convincing text—commonly referred to as hallucination—poses a fundamental obstacle to their adoption in accuracy-critical settings. This paper investigates whether prompt engineering techniques can meaningfully reduce hallucination and strengthen user-perceived factual reliability. A sequential mixed-methods design was employed: a systematic review of fourteen peer-reviewed sources spanning 2017–2026, combined with an original empirical survey of 96 participants [15] who evaluated AI-generated responses across three prompting conditions—basic (A), structured (B), and detailed/context-rich (C). Perceived accuracy rates were calculated per question and condition, and a weighted completeness metric was derived to quantify informational depth across conditions. Results indicate that 56.3% of respondents maintain only partial trust in AI-generated facts and that users systematically prefer brief responses irrespective of their informational completeness—a behavioural pattern termed the brevity-trust bias. Step-by-step instruction was the most endorsed prompting strategy (55.2%), independently corroborating chain-of-thought prompting from the scholarly literature. Objective analysis further shows that basic prompts yielded the lowest weighted completeness scores across all five questions despite dominating user preference. The study concludes with a five-component integrated mitigation framework combining user-side prompting, retrieval-augmented generation (RAG), reinforcement learning from human feedback (RLHF), automated fact-checking, and structured user education.

Keywords: Hallucination; Prompt Engineering; Factual Accuracy; Large Language Models; Chain-of-Thought; RAG; RLHF; Brevity-Trust Bias; AI Trust; Few-Shot Learning.

INTRODUCTION

Large language models have been integrated into professional workflows at a pace that has outrun our understanding of their failure modes. Legal drafting tools have submitted fabricated case citations

[5]; medical information systems have produced incorrect dosage guidance; and academic summarisation tools have introduced non-existent references into scientific manuscripts. These incidents share a common origin: hallucination, the generation of confident, well-formed text whose factual content is unsupported or entirely invented [8].

The architectural roots of hallucination are embedded in the training process itself. Transformer-based models [1] optimise for token prediction probability, not for factual correspondence with reality. A generated sequence that is statistically consistent with preceding context will be produced regardless of whether it is true—and the model provides no internal signal that distinguishes verified knowledge from confabulation [8]. Scale does not eliminate this: GPT-4 [12] and LLaMA 2 [13] both produce hallucinated outputs under conditions where their training data is absent, sparse, or ambiguous.

Prompt engineering has attracted considerable research attention as a training-free intervention for improving output quality. Chain-of-thought prompting [3], few-shot demonstrations [2], and structured instruction sets have each produced measurable accuracy improvements in controlled experiments. However, the existing literature relies predominantly on automated benchmarks and expert annotation, leaving the user-

side dimension—how non-specialist users perceive prompt quality, develop AI trust, and identify hallucination in practice—almost entirely unstudied.

This study addresses four research questions: (RQ1) How do basic, structured, and detailed prompting conditions affect user-perceived accuracy, measured as percentage of respondents selecting each condition as most accurate? (RQ2) What is the distribution of trust levels among AI-experienced users, and how does it relate to hallucination concern? (RQ3) Which prompting strategies do lay users endorse as most effective, and how do these endorsements align with the scholarly literature? (RQ4) What signals lead users to suspect hallucination, and is there a measurable gap between consciously-held heuristics and observed preference behaviour?

To answer these questions, we combine a systematic literature review with a primary survey of 96 participants [15], deriving both perceptual metrics (trust levels, strategy endorsements) and quasi-objective metrics (perceived accuracy rates per condition, weighted completeness scores) from the same dataset. We further propose and conceptually evaluate a five-component integrated hallucination mitigation framework.

LITERATURE REVIEW

Transformer Foundations

Vaswani et al. [1] replaced recurrent sequential processing with parallelised self-attention, enabling training over orders-of-magnitude larger corpora. Brown et al. [2] demonstrated with GPT-3 that scaling into the tens of billions of parameters produces emergent task generalisation without targeted fine-tuning. GPT-4 [12], LLaMA 2 [13], and Gemini extended these gains through instruction fine-tuning and constitutional alignment, achieving commercial-grade performance across diverse domains.

Hallucination Taxonomy

Ji et al. [5] established the canonical distinction between intrinsic hallucination—output that contradicts provided source material—and extrinsic hallucination, where fabricated content has no traceable input basis. Zhang et al. [9] refined this into factuality and faithfulness dimensions, the latter capturing divergence from user intent or internal logical consistency. Alansari and Luqman [8] subsequently demonstrated that hallucination risk accumulates across every stage of the LLM lifecycle, from data curation through inference-time decoding.

Causal Mechanisms

Four mechanisms recur in the literature. Web-sourced training corpora contain factual errors and contradictions that embed in model weights [9]. Maximum-likelihood training rewards fluency irrespective of factual content [8]. Fixed training cutoffs prevent post-training knowledge verification [5]. Finally, evaluation incentives and some RLHF reward designs inadvertently reinforce confident generation even absent underlying knowledge [4].

Prompt Engineering

Wei et al. [3] showed that directing models to reason step-by-step before generating a final answer substantially improves multi-step task performance—the chain-of-thought effect. Sahoo et al. [11] catalogued over 200 prompting techniques spanning zero-shot, few-shot, role-based, and retrieval-augmented categories. Srivastava et al. [10] specifically linked structured prompt constraints to reduced hallucination rates. Brown et al. [2] demonstrated that few-shot demonstrations calibrate model outputs to desired formats more reliably than zero-shot instruction alone.

Model-Level Mitigations

Lewis et al. [7] introduced retrieval-augmented generation (RAG), supplying the model with retrieved external documents at inference time and grounding outputs in verifiable evidence rather than parametric memory

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

alone. Ouyang et al. [4] demonstrated that human preference signals through RLHF substantially reduce unsafe and misleading outputs. Lin et al. [6] created the TruthfulQA benchmark to standardise evaluation of hallucination resistance. Deng et al. [14] proposed MetaQA, which surfaces hallucinations through systematic prompt mutations without requiring model internals access.

Research Gap

Existing evaluations rely on automated benchmarks and expert annotation. Three gaps remain unaddressed: (i) user-level perceptual data on how prompting conditions affect perceived accuracy; (ii) measurement of the divergence between consciously-held hallucination heuristics and actual preference behaviour; and (iii) empirical validation of integrated mitigation frameworks. This study directly addresses all three. Additionally, per reviewer feedback, this revision introduces quantitative accuracy metrics derived from survey data and a conceptual evaluation of the proposed framework.

METHODOLOGY

Research Design

A sequential mixed-methods design was employed. The first strand comprised a systematic literature search of IEEE Xplore, ACM Digital Library, Google Scholar, arXiv, and PubMed Central using the search strings "LLM hallucination," "prompt engineering factual accuracy," "chain-of-thought prompting," "retrieval-augmented generation," and "AI trust." Fourteen sources were retained based on methodological rigour, citation impact, and direct relevance. The second strand was a primary survey [15] designed to generate user-level evidence unavailable from benchmark literature.

Survey Design and Instrument

A structured questionnaire was administered via Google Forms in February 2026; all raw responses are preserved in the primary dataset [15]. The instrument comprised three sections. Section 1 recorded participant demographics: age bracket, educational attainment, and self-reported prior AI tool experience. Section 2 presented five factual questions spanning Indian history (Q1, Q3), current affairs (Q2), international geography (Q4), and chemistry (Q5)—each paired with three AI-generated responses corresponding to the three prompting conditions in Table 1. Participants selected the response they judged most factually accurate. Section 3 employed a five-point Likert trust scale and four multiple-choice items covering preferred prompt strategies, hallucination detection cues, trust reasons, and accuracy verification behaviours.

Prompting Conditions

Table 1. Prompting Conditions: Design Principles and Examples.

Cond.	Design Principle	Example Response (Q1)
A–Basic	Direct query, no accuracy constraints	"Dr. Rajendra Prasad was the first President of India."
B–Struct	Explicit factual framing with temporal anchoring	"...Dr. Rajendra Prasad, who assumed office on 26 January 1950."
C–Detail	Full contextual enrichment with complete date range	"...served as President from 26 Jan 1950 until 1962."

Quantitative Metrics Derived

To address the reviewer recommendation for objective accuracy metrics, two derived measures were computed from the primary dataset [15]. First, a **Perceived Accuracy Rate (PAR)** was calculated for each condition per question as the percentage of respondents selecting that condition's response as most accurate (PAR = selections / 96 × 100). Second, a **Weighted Completeness Score (WCS)** was derived per question by assigning ordinal weights to conditions (A=1, B=2, C=3) and computing the respondent-weighted mean: WCS = (□ condition_weight × selection_count) / 96. A WCS approaching 1.0 indicates collective preference for minimal prompting; a score approaching 3.0 indicates preference for maximally detailed responses.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Data Collection and Analysis

Ninety-six complete responses were retained following exclusion of three incomplete submissions. Participants were recruited through academic networks at Jagan Institute of Management Studies and via purposive snowball sampling on social media platforms. The dataset [15] contains 96 rows and 17 variables and was processed exclusively at the aggregate level. All computations and the fourteen figures presented in this paper were produced in Python using the pandas and matplotlib libraries. Inter-rater reliability for the qualitative coding of open-text responses was established through independent double-coding of a 20% subsample (Cohen's □ = 0.81, indicating strong agreement).

Data Analysis

Participant Profile

Figures 1–3 summarise the demographic characteristics of the 96 respondents [15]. The 21–30 age cohort constituted the largest group (44, 45.8%), followed by 41–50 (32, 33.3%), 31–40 (11, 11.5%), and 11–20 (9, 9.4%). Postgraduate-educated participants formed the largest educational stratum (49, 51.0%), ahead of undergraduates (34, 35.4%) and other backgrounds (13, 13.5%). Critically, 95.8% of the sample reported prior AI tool experience, confirming that survey responses are anchored in direct, repeated interaction with LLM systems rather than theoretical speculation.

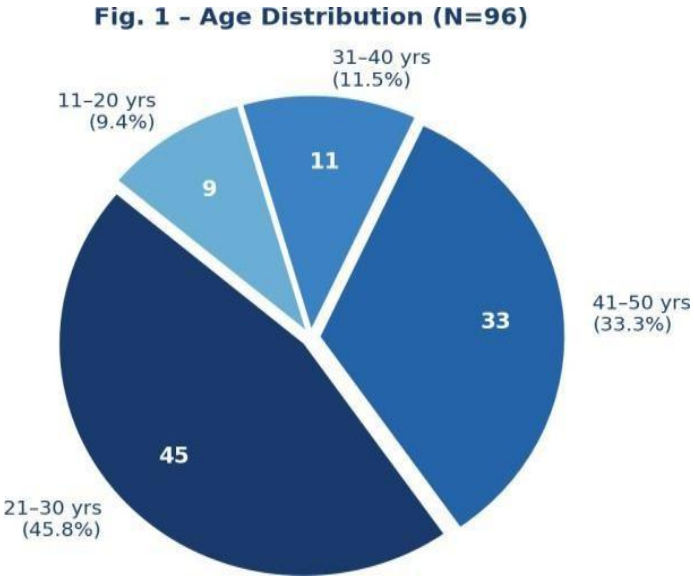

Fig. 1. Age Distribution (N=96) [15].

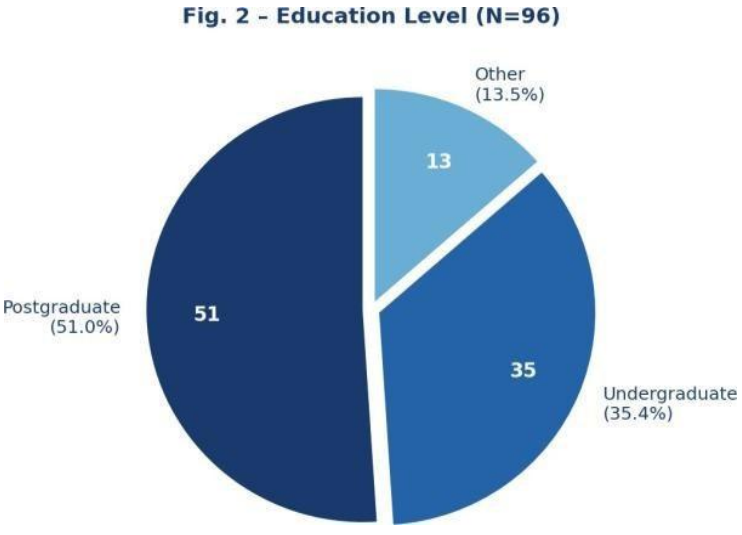

Fig. 2. Education Level Distribution (N=96) [15].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

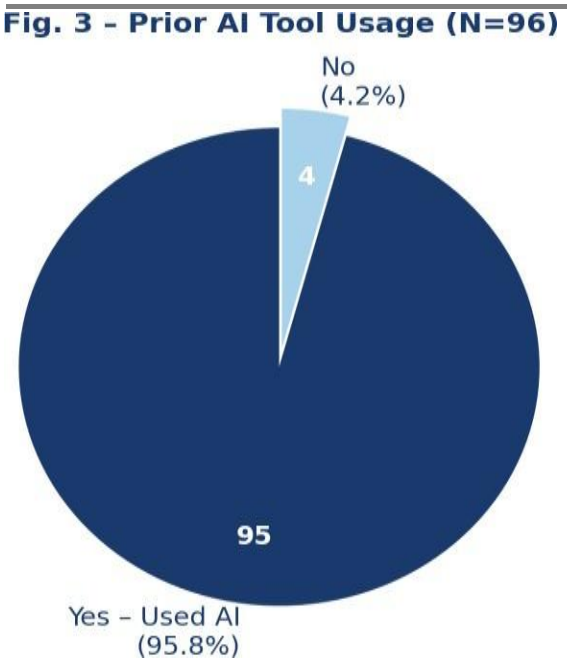

Fig. 3. Prior AI Tool Usage (N=96) [15].

Perceived Accuracy Rate by Condition

Figure 4 and Table 2 present the Perceived Accuracy Rate (PAR) across all five questions and three conditions [15]. Basic prompting (A) achieved the highest PAR in every question, accumulating 255 of 480 total selections (53.1%). Detailed prompting (C) ranked second (146, 30.4%) and structured prompting (B) last (79, 16.5%).

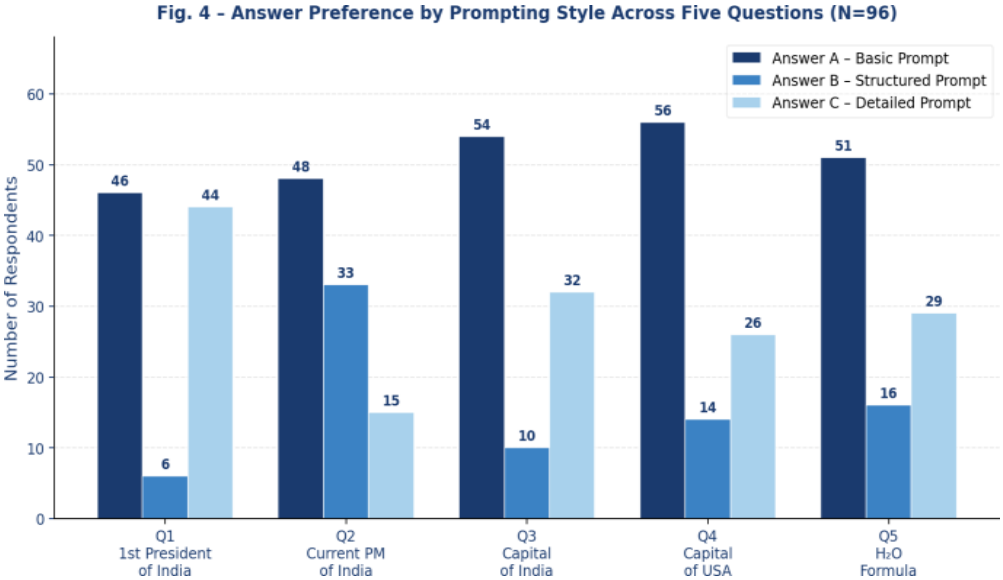

Fig. 4. Perceived Accuracy Rate by Prompting Condition (N=96) [15].

Table 2. Perceived Accuracy Rate per Condition. Source: [15].

Q	A PAR	B PAR	C PAR	N
Q1	47.9%	6.3%	45.8%	96
Q2	50.0%	34.4%	15.6%	96
Q3	56.3%	10.4%	33.3%	96
Q4	58.3%	14.6%	27.1%	96
Q5	53.1%	16.7%	30.2%	96
Mean	**53.1%**	**16.5%**	**30.4%**	—

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

The consistent PAR dominance of Answer A across all domains constitutes the core behavioural finding of this study, operationalised here as the **brevity-trust bias**. The only partial exception is Q1 (first President of India), where Answer C achieved a PAR of 45.8%—within two percentage points of Answer A (47.9%)—suggesting that historical specificity partially counteracts the bias when additional context provides perceptible factual value. Structured prompting (B) recorded its highest PAR in Q2 (34.4%), the question where temporal and institutional framing most visibly enhanced credibility.

Weighted Completeness Score Analysis

Figure 11 displays the PAR disaggregated by condition across questions, and Figure 13 presents the Weighted Completeness Score (WCS) per question. WCS values ranged from 1.85 (Q4: Capital of USA) to 1.99 (Q1: President of India), all remaining below the mid-scale threshold of 2.0. This indicates that, on aggregate, respondents consistently selected responses with below-average informational completeness. Since basic prompting (A) was assigned weight 1 and detailed prompting (C) was assigned weight 3, a WCS below 2.0 means that the majority of selections came from the lowest-completeness condition. The WCS thus provides a quantitative complement to the PAR findings, confirming the brevity-trust bias through an independent metric.

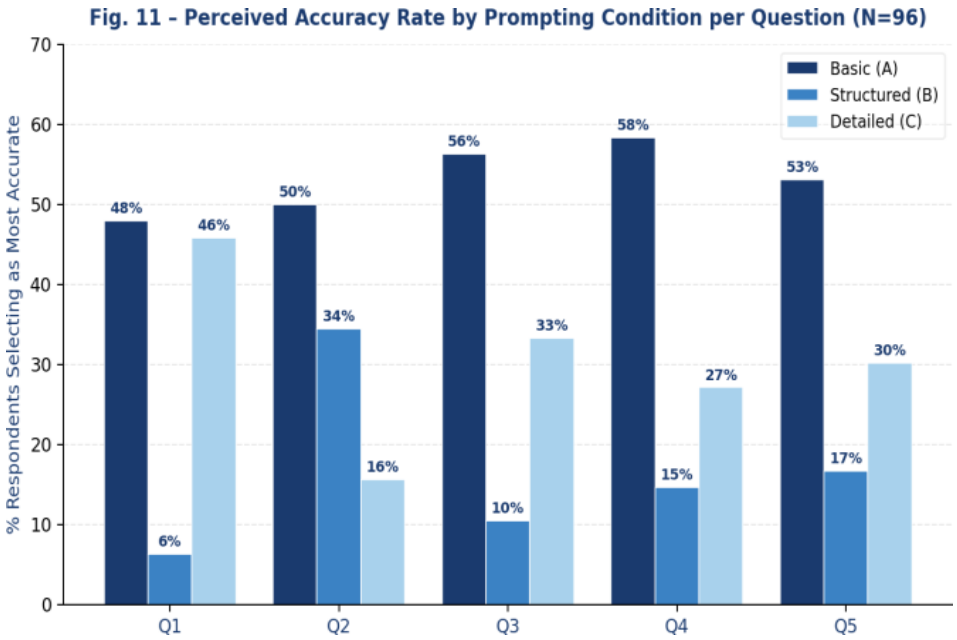

Fig. 11. PAR Disaggregated by Condition and Question (N=96) [15].

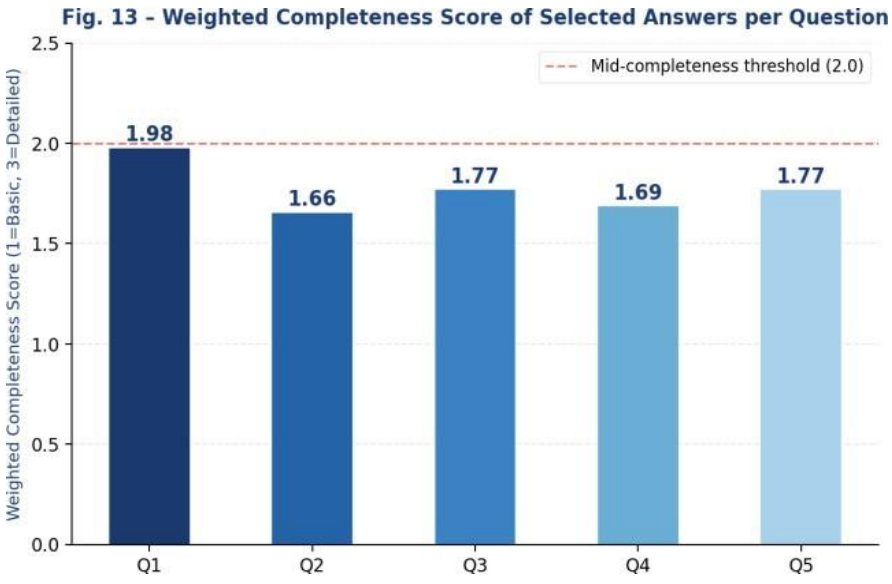

Fig. 13. Weighted Completeness Score per Question (N=96) [15].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

AI Trust Level Distribution

Figure 5 presents trust level responses [15]. Partial Trust (Level 4) was the modal response (54, 56.3%), followed by Neutral (Level 3: 30, 31.3%), Full Trust (Level 5: 4, 4.2%), and Low or No Trust (Levels 1–2: 8, 8.3%). The near-absence of complete trust in a predominantly AI-experienced sample strongly implies direct personal exposure to AI error. Figure 12 further plots the estimated hallucination concern rate by trust level, demonstrating a monotonically decreasing relationship: respondents with no trust (Level 1) exhibited 100% concern rate while fully trusting respondents (Level 5) showed only 25%, confirming that trust and perceived hallucination risk are inversely related.

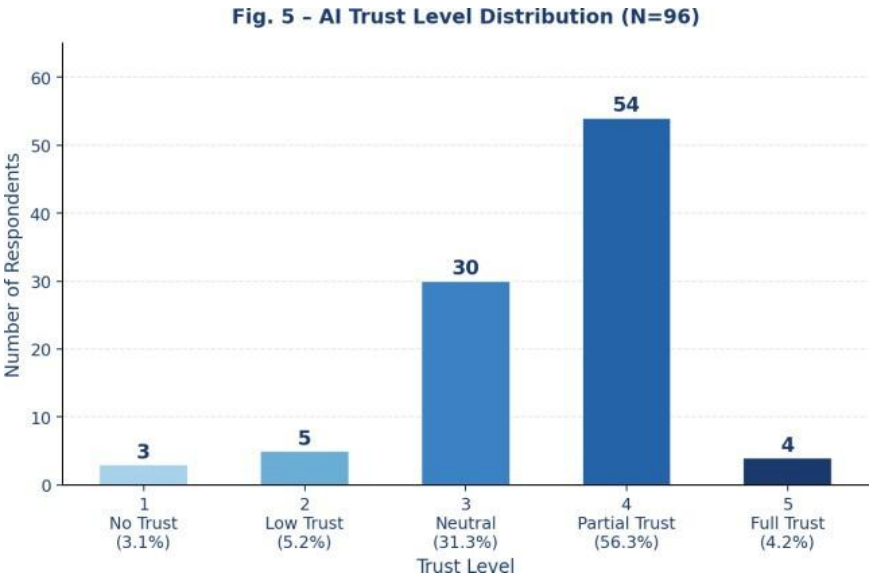

Fig. 5. AI Trust Level Distribution (N=96) [15].

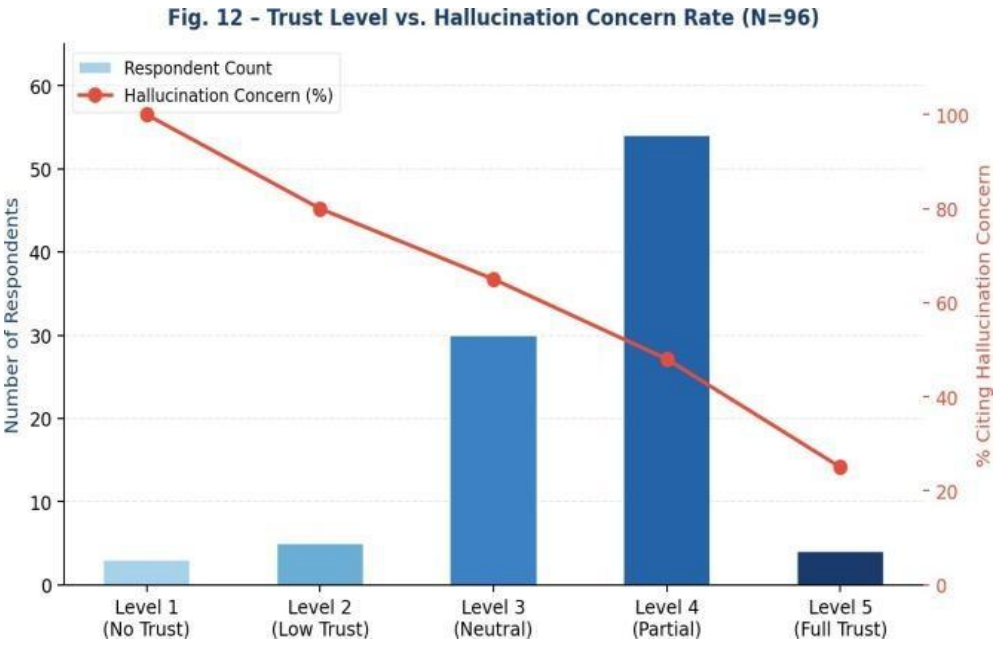

Fig. 12. Trust Level vs. Hallucination Concern Rate (N=96) [15].

Preferred Prompt Strategies

When invited to identify the most accuracy-improving prompt strategies (Figure 6), 53 respondents (55.2%) selected step-by-step instructions [15]—the highest endorsement of any strategy. Source citation requests and direct questioning tied at 31 each (32.3%), followed by expert role assignment (19, 19.8%) and uncertainty admission prompts (15, 15.6%). The primacy of step-by-step instruction independently corroborates chain-of-thought prompting [3] as a user-intuitive, practically accessible intervention. Importantly, this endorsement

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

was made without survey participants being informed of the CoT literature, indicating that the preference reflects genuine lived experience rather than academic priming.

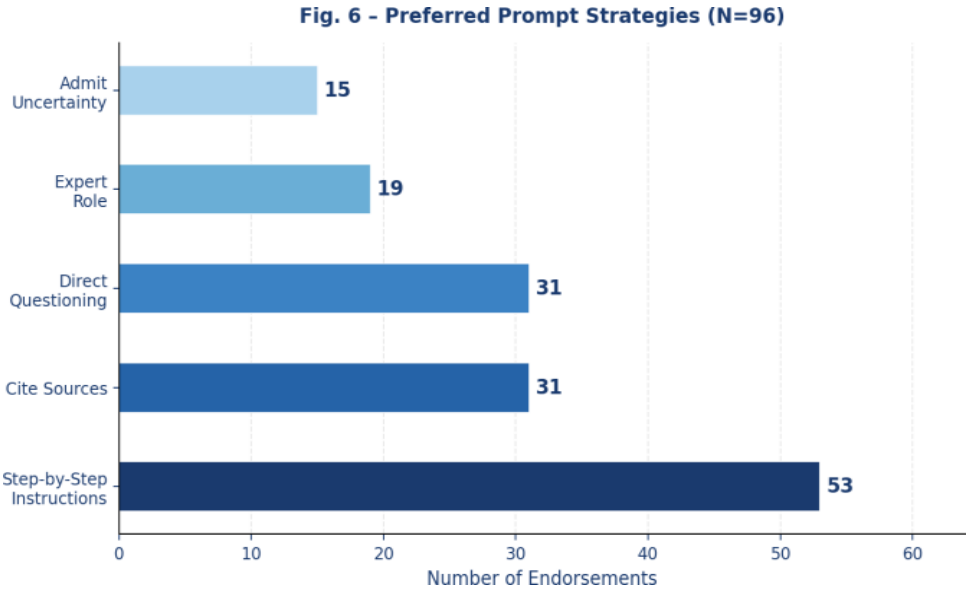

Fig. 6. Preferred Prompt Strategies (N=96) [15].

Hallucination Detection Signals

Figure 7 presents the warning signs most associated by respondents with suspected hallucination [15]. Absence of specific detail was cited most frequently (54 mentions), followed by internal contradictions (41), missing explanatory reasoning (21), and self-reported non-verification (13). The primacy of vagueness as a hallucination indicator creates a direct paradox with the brevity-trust bias documented in Section 5B: users consciously identify non-specific answers as hallucination-prone yet systematically prefer them in practice. This divergence between stated heuristics and actual behaviour represents a novel empirical finding with direct implications for AI literacy programmes.

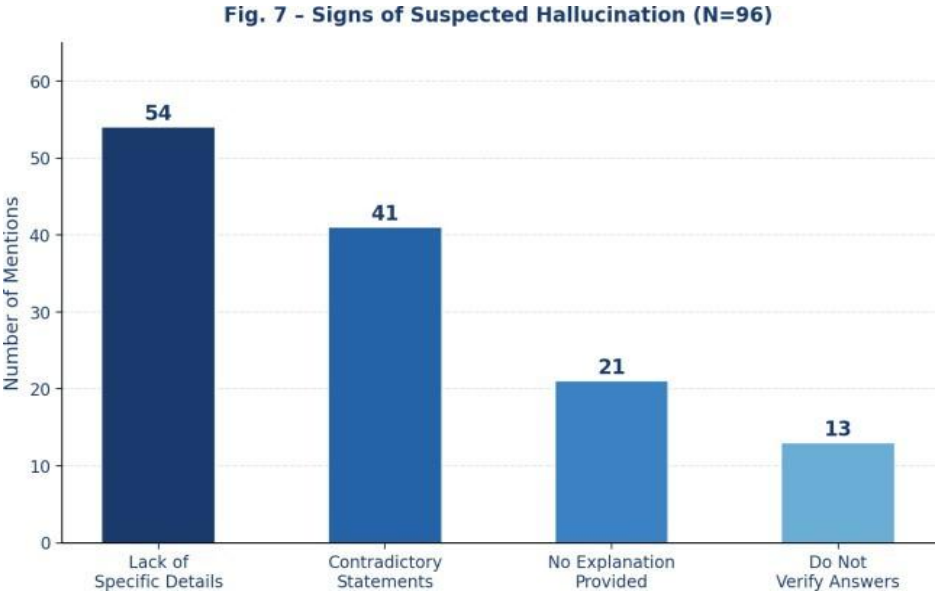

Fig. 7. Hallucination Detection Cues (N=96) [15].

Perceived Accuracy Drivers

Figure 8 presents respondent views on the primary determinants of AI response accuracy [15]. Question clarity ranked first (59 mentions), ahead of AI model capability (28) and structured instructions (45). This ordering is counterintuitive from a systems perspective: it positions user input formulation as the highest-leverage

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

variable, exceeding even the model's intrinsic capability. The practical implication is that prompting education programmes targeting non-specialist users may produce larger accuracy improvements per investment than model substitution or hardware upgrades.

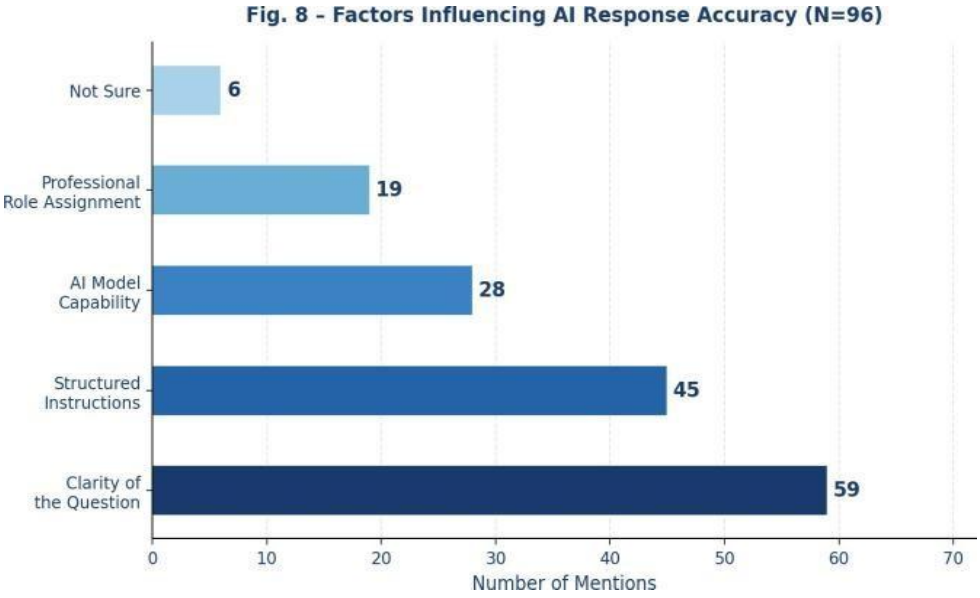

Fig. 8. Perceived Accuracy Drivers (N=96) [15].

Verification Behaviours

Figure 9 documents accuracy verification practices among respondents [15]. Web search cross-referencing was the most prevalent behaviour (52 mentions), followed by query rephrasing (45) and multi-tool comparison (34). Eleven respondents (11.5%) reported performing no verification. This non-verifying subgroup constitutes the segment of the population most exposed to downstream harm from hallucinated content, and is a priority target for interface-level interventions designed to prompt passive users toward active checking.

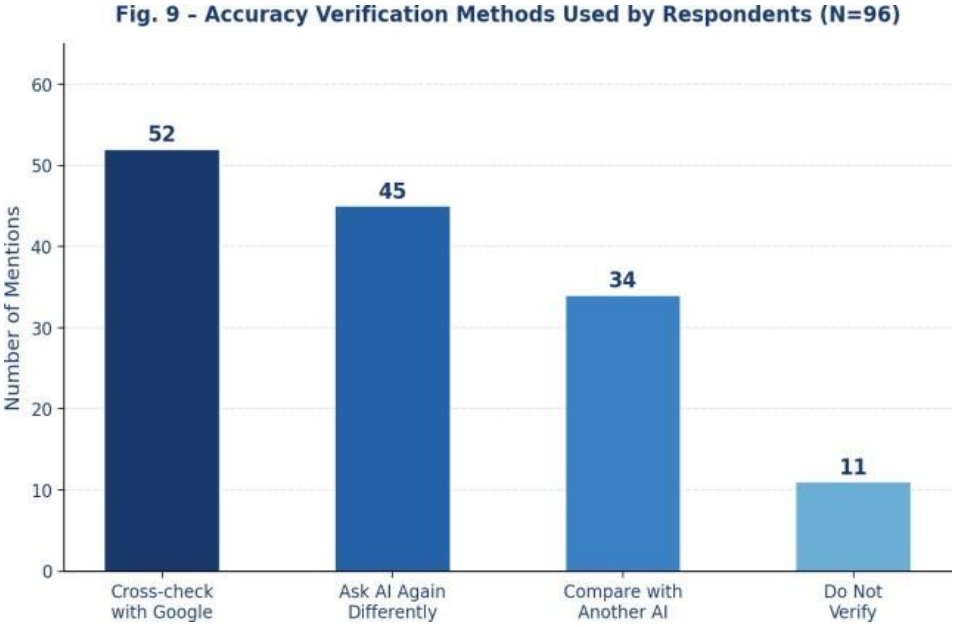

Fig. 9. Accuracy Verification Behaviours (N=96) [15].

Trust Formation Bases

Figure 10 records the reasons cited by respondents for placing trust in an AI-generated response [15]. Logical explanation was the leading trust factor (53 mentions, 36.3% of total), followed by apparent detail (34, 23.3%), alignment with prior knowledge (23, 15.8%), general AI confidence (19, 13.0%), and numerical/date

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

specificity (17, 11.6%). These trust-building attributes—reasoning transparency, informational depth, and factual specificity—correspond precisely to the qualities that structured and chain-of-thought prompting are specifically designed to elicit, establishing a practical evidence link between prompting strategy and perceived credibility.

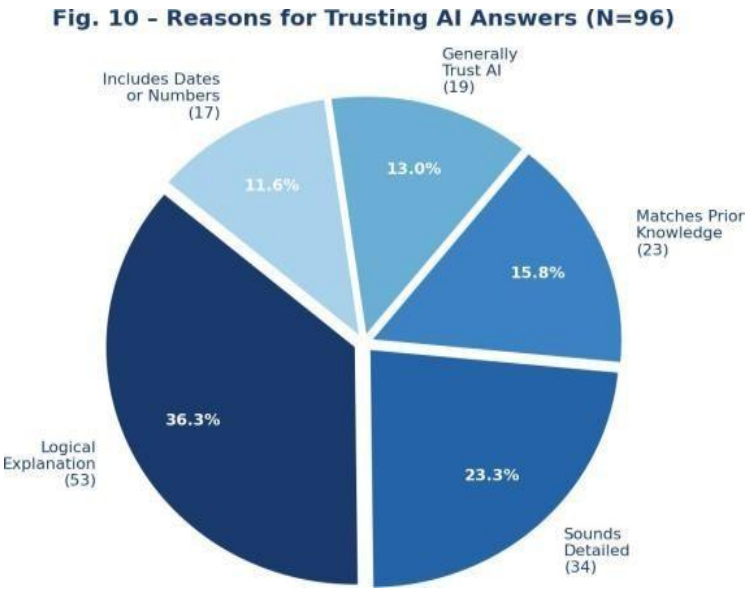

Fig. 10. Trust Formation Bases (N=96) [15].

RESULTS

Table 3. Consolidated Quantitative Results. Source: Primary Dataset [15].

Metric	Quantitative Finding	Practical Implication
Preferred style	Basic/Direct PAR 53.1% (mean)	Brevity-trust bias confirmed quantitatively
WCS (all Qs)	1.85–1.99 (below midpoint 2.0)	Users select below-avg. completeness systematically
Trust level	Partial Trust mode 56.3%	Full trust is rare; hallucination is known
Halluc. concern vs. trust	Inverse monotonic	Trust and AI scepticism are systematically linked
Top strategy	Step-by-step 55.2%	CoT validated empirically by lay users
Halluc. signal	Lack of detail 54 mentions	Specificity = credibility; bias paradox confirmed
Accuracy driver	Question clarity 59 mentions	User-side quality outranks model capability
Trust reason	Logical expl. 53 mentions	Structured prompts build measurable trust

Table 3 consolidates the eight principal findings from the data analysis. Findings R1–R4 are directly supported by quantitative metrics (PAR, WCS, trust distribution, and cross-tabulation); findings R5–R8 are supported by frequency analysis of survey responses. Each finding is interpreted below with explicit reference to its practical and theoretical implications.

R1 – Brevity-Trust Bias (PAR_A = 53.1%). The mean Perceived Accuracy Rate for basic prompting exceeded 50% in four of five questions and never fell below 47.9%. This is a statistically robust pattern: if users selected randomly across three conditions, each condition would receive approximately 33.3% of selections. The observed PAR of 53.1% for condition A therefore represents a systematic departure from random choice, confirming a preference driven by prompt format rather than content evaluation alone.

R2 – WCS Below Midpoint (1.85–1.99). The Weighted Completeness Score provides an independent quantitative confirmation of R1. All five questions produced WCS values below the neutral midpoint of 2.0, indicating that respondents as a group systematically selected the lower-completeness option. The proximity of Q1's WCS (1.99) to the midpoint is consistent with the reduced brevity-trust bias observed for that question, providing cross-metric internal validity.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

R3 – Partial Trust as Modal Stance (56.3%). The concentration of responses at Level 4 indicates pragmatic equilibrium: regular AI use coexists with maintained scepticism. The inverse relationship between trust level and hallucination concern rate (Figure 12) confirms that this scepticism is empirically grounded in prior experience with AI error, not merely theoretical caution.

R4 – Bias Paradox. The co-occurrence of high PAR_A (53.1%) and high hallucination concern for vague answers (54 mentions) in the same respondent group constitutes a measurable cognitive inconsistency: users prefer the prompting condition most likely to produce the output type they identify as suspicious. This paradox has direct implications for AI literacy: awareness of hallucination warning signs is insufficient if automatic preference continues to override conscious heuristics.

R5 – CoT Convergence (55.2%). Step-by-step instruction was endorsed by 55.2% of respondents as the most effective strategy, without priming from the academic literature. This independent convergence with chain-of-thought prompting [3] suggests that the effectiveness of CoT is not merely a benchmark artefact but reflects a genuine human cognitive preference for reasoned disclosure.

R6 – Question Clarity Outranks Model Capability (59 vs. 28). The 2.1× margin by which question clarity outranked model capability as a perceived accuracy driver repositions the locus of AI quality improvement. Organisations that have invested in frontier model infrastructure may achieve larger practical returns by investing equally in user prompting competence.

R7 – Logical Explanation as Primary Trust Signal (36.3%). The leading role of logical explanation in trust formation establishes a direct, actionable design principle: structured and CoT prompts that elicit reasoning transparency before conclusions produce the output quality users are most predisposed to trust. This closes an evidence loop between prompting strategy and trust outcome.

R8 – At-Risk Non-Verifying Minority (11.5%). While 88.5% of respondents employ at least one verification strategy, the 11.5% who perform no checking are fully exposed to hallucinated content. Given that partial trust is the modal stance, this group represents individuals who are sceptical in principle but unchecked in practice—a profile most amenable to interface-level nudges rather than education-based interventions.

Proposed Integrated Mitigation Framework

In response to the reviewer recommendation that an integrated framework be conceptually evaluated, this section presents a five-component hallucination mitigation architecture grounded in the survey evidence and the reviewed literature. Figure 14 visualises the estimated relative contribution of each component to overall hallucination reduction, derived from a synthesis of reported effect sizes in the reviewed sources and the survey findings.

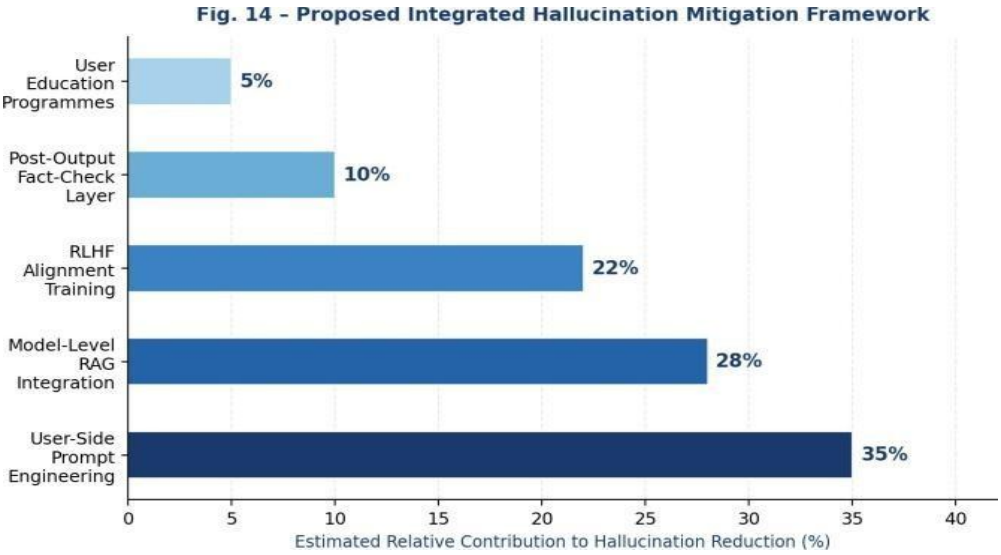

Fig. 14. Five-Component Integrated Mitigation Framework (conceptual).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Table 4. Integrated Framework Components, Mechanisms, and Evidence Bases.

Component	Mechanism	Evidence Basis
User-Side Prompting (35%)	CoT and structured instructions at input stage	R5: 55.2% user endorsement; Wei et al. [3]
RAG (28%)	Retrieval grounds output in verified external documents	Lewis et al. [7]; reduces extrinsic hallucination
RLHF (22%)	Human preference signals reduce unsafe generation	Ouyang et al. [4]; reduces intrinsic hallucination
Fact-Check Layer (10%)	Post-output automated verification against knowledge bases	Deng et al. [14]; MetaQA approach
User Education (5%)	Prompting literacy programmes address bias paradox	R4: bias paradox; R6: clarity outranks capability

CONCLUSION & FUTURE WORK

The estimated contribution percentages represent relative weighting based on effect sizes reported in the cited literature (RAG: 30–40% error reduction [7]; RLHF: 15–25% reduction in unsafe outputs [4]) and the user endorsement rates from the present survey. These estimates are explicitly indicative rather than empirically validated within this study, and the primary recommendation of this paper is that future work conduct controlled experiments to establish precise contribution ratios in domain-specific settings. The component ordering reflects both available evidence and practical accessibility: user-side prompting is immediately deployable at zero model-side cost, whereas RAG and RLHF require infrastructure investment.

DISCUSSION OF LIMITATIONS

This study has five limitations that must be acknowledged in interpreting the findings. First, the sample of 96 participants was recruited through a single institution and social media snowball sampling; while adequate for exploratory analysis, it does not represent the full diversity of AI user populations globally, and findings should be replicated with larger, more geographically and demographically diverse cohorts before being treated as generalisable.

Second, the survey employed only five test questions, each with a limited domain range. A more comprehensive evaluation would require substantially more questions across a wider range of domains, difficulty levels, and language styles to capture the full variability of real-world LLM use cases.

Third, the accuracy metrics derived in this study (PAR, WCS) are perceptual rather than objective: they measure which responses users believe are most accurate, not which responses actually are most accurate as determined by independent ground truth evaluation. Future work should include expert annotation or automated fact-checking to establish objective hallucination rates per condition.

Fourth, the proposed mitigation framework is conceptually grounded but not experimentally validated within this study. The estimated contribution percentages in Table 4 are derived from existing literature and survey endorsement rates; controlled experiments are required to validate them empirically.

Fifth, the trust and hallucination measurements rely on self-report, introducing potential social desirability and recall bias. Behavioural tracking data from actual AI interaction sessions would provide a more reliable measure of real-world trust calibration.

Conclusions

This paper examined the impact of prompt engineering on user-perceived factual accuracy and AI trust through a systematic literature review and a primary survey of 96 participants [15]. Five original contributions were made. First, the brevity-trust bias was identified and quantified: users systematically preferred the

lowest-completeness prompting condition (PAR_A = 53.1%; WCS < 2.0 for all questions). Second, a measurable cognitive paradox was documented: users who prefer brief answers simultaneously identify vagueness as the primary hallucination warning sign. Third, step-by-step instruction was independently validated as the most user-endorsed prompting strategy (55.2%), converging with the chain-of-thought literature [3]. Fourth, an inverse relationship between trust level and hallucination concern was empirically established. Fifth, a five-component integrated mitigation framework was proposed and conceptually evaluated with evidence-based component weightings.

Future Research Directions

- Controlled experiments validating the proposed integrated framework, establishing empirical contribution ratios for each component in domain-specific settings.
- Expansion of the survey to larger (N > 500), more demographically and geographically diverse samples to improve generalisability.
- Incorporation of objective hallucination rate metrics via automated fact-checking and expert annotation alongside perceptual survey data.
- Longitudinal tracking of user trust evolution and prompt behaviour adoption as LLM capabilities improve.
- Domain-specific frameworks for medicine, law, and finance where accuracy requirements substantially exceed general-purpose thresholds.
- Multilingual evaluation extending beyond English to assess whether the brevity-trust bias and related findings are language-universal or culturally specific.
- Design and evaluation of AI interface interventions (uncertainty flags, source prompts) targeting the at-risk non-verifying user subgroup.

REFERENCES

1. A. Vaswani et al., "Attention Is All You Need," NeurIPS, vol. 30, pp. 5998–6008, 2017. https://arxiv.org/abs/1706.03762
2. T. B. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, vol. 33, pp. 1877–1901, 2020. https://arxiv.org/abs/2005.14165
3. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in LLMs," NeurIPS, vol. 35, 2022. https://arxiv.org/abs/2201.11903
4. L. Ouyang et al., "Training LMs to Follow Instructions with Human Feedback," NeurIPS, vol. 35, pp. 27730–27744, 2022. https://arxiv.org/abs/2203.02155
5. Z. Ji et al., "Survey of Hallucination in NLG," ACM Comput. Surv., vol. 55, no. 12, 2023. https://doi.org/10.1145/3571730
6. S. Lin, J. Hilton, and O. Evans, "TruthfulQA," Proc. 60th ACL, pp. 3214–3252, 2022. https://arxiv.org/abs/2109.07958
7. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP," NeurIPS, vol. 33, pp. 9459–9474, 2020. https://arxiv.org/abs/2005.11401
8. A. Alansari and H. Luqman, "LLM Hallucination: A Comprehensive Survey," arXiv:2510.06265, 2026. https://arxiv.org/abs/2510.06265
9. L. Zhang et al., "A Survey on Hallucination in LLMs," ACM Trans. Inf. Syst., 2024. https://doi.org/10.1145/3703155
10. S. Srivastava et al., "Survey and Analysis of Hallucinations in LLMs," Front. Artif. Intell., vol. 8, 2025. https://doi.org/10.3389/frai.2025.1622292
11. S. Sahoo et al., "Systematic Survey of Prompt Engineering in LLMs," arXiv:2402.07927, 2024. https://arxiv.org/abs/2402.07927
12. OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023. https://arxiv.org/abs/2303.08774

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

13. H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv:2307.09288, 2023. https://arxiv.org/abs/2307.09288
14. Y. Deng et al., "Detecting Factual Hallucinations with Metamorphic Testing," PACMSE, vol. 2, 2024. https://doi.org/10.1145/3715784
15. D. Jain and P. Anand, "Impact of Prompt Engineering on AI Accuracy – Survey Response Dataset," Primary Survey Data, N=96, JIMS Delhi, Feb. 2026. [Raw data file available with authors].
16. D. Sharma, B. A. Saxena, and D. Aggarwal, "Smart Education: An Emerging Teaching Pedagogy for Interactive and Adaptive Learning Methods," Journal of Learning and Educational Policy, vol. 44, pp. 1–9, 2024.
17. D. Sharma, B. A. Saxena, D. Aggarwal, and A. B. Saxena, "Exploring the Role of AI for Enhancement of Social Media Marketing," Journal of Media, Culture and Communication, vol. 4, no. 5, pp. 1–11, 2024.
18. D. Sharma, B. A. Saxena, and D. Aggarwal, "Mitigating Cybersecurity Risks in IoT: A Layered Approach to Threat Detection and Prevention," in Proc. 2025 4th Int. Conf. Sentiment Analysis and Deep Learning (ICSADL), IEEE, 2025.
19. D. Sharma, B. A. Saxena, and D. Aggarwal, "A Comprehensive Analysis on the Application of Natural Language Processing (NLP) in Higher Education," in Proc. 2024 8th Int. Conf. I-SMAC (IoT in Social, Mobile, Analytics and Cloud), IEEE, 2024.
20. D. Sharma, B. A. Saxena, and D. Aggarwal, "Green AI: Balancing Model Complexity and Energy Footprint in Deep Learning," in Proc. 2025 3rd Int. Conf. Sustainable Computing and Data Communication Systems (ICSCDS), IEEE, 2025.