Intent Hallucination in Large Language Models: A Systematic Review of Instruction Following, Detection, And Mitigation Techniques

Sai Gadhave[1], Krutika Sonare[2], Anushka Varpe[3], Anushka Waghchoure[4], Aaryan Laygude[5], Dr. Shital Ghotekar[6]

[1,2,3,4,5,6]*MAEER's MIT Arts, Commerce and Science College Alandi(D), Pune*

doi.org/10.64643/ijirt.208736-459

***Abstract*—Large Language Models (LLMs) have demonstrated significant capabilities in generating human-like and contextually relevant responses across a wide range of applications. However, their reliability is affected not only by factual hallucinations but also by failures to correctly understand and follow user instructions. This emerging issue, referred to as intent hallucination, occurs when an LLM omits, misinterprets, or fails to satisfy one or more requirements contained in a user's request, even when the generated response may be factually correct. Such failures can reduce the reliability and usefulness of LLM-based systems, particularly in complex tasks involving multiple constraints.**

This paper proposes a systematic review of existing research on intent hallucination and instruction-following capabilities in Large Language Models. The study will examine major types and causes of instruction-following failures, existing benchmarks and evaluation techniques, methods for detecting such failures, and approaches proposed to mitigate them. The review will also compare current evaluation frameworks and investigate limitations in existing research, particularly concerning complex, multi-constraint, and dynamic user instructions. Based on the reviewed literature, the study aims to organize existing approaches into a structured taxonomy covering instruction-following failures, detection and evaluation methods, and mitigation strategies. Finally, the paper will highlight open research challenges and future directions for developing more reliable and intent-aligned LLM systems. The study aims to provide a consolidated understanding of intent hallucination and support future research toward more dependable human–AI interaction.

***Index Terms*—Large Language Models, Generative AI, Intent Hallucination, Instruction Following, LLM Evaluation, Hallucination Detection, Prompt Engineering, Artificial Intelligence**

I. INTRODUCTION

Large Language Models (LLMs) are increasingly used for writing, programming, tutoring, customer support, data analysis, and decision support. Their usefulness depends not only on producing factually plausible text, but also on accurately carrying out the user's intended request. A response can be factually correct yet still be unsatisfactory if it ignores a required format, omits a requested section, uses the wrong language, violates a word limit, or fails to follow several instructions simultaneously.

This problem can be described as intent hallucination: a mismatch between the user's expressed requirements and the model's generated output. Unlike factual hallucination, where a model produces unsupported or false information, intent hallucination occurs when the model misunderstands, overlooks, reinterprets, or incompletely executes an instruction. For example, a user may request "a 200-word summary in five bullet points without technical terms," while the model produces a well-written summary that exceeds the limit, uses paragraphs, and includes specialist terminology. The content may be accurate, but the response has not fulfilled the task.

Instruction-following failures are particularly important in complex prompts. Many real-world requests combine content requirements, output formats, style constraints, exclusions, priorities, and interaction rules. A model must identify each requirement, determine whether requirements conflict, preserve constraints across turns, and verify that its final answer satisfies all relevant conditions. Performance often declines as instructions become

more numerous, nuanced, or dynamically updated during a conversation.

Recent benchmarks have begun to evaluate this capability directly. IFEval assesses whether models follow objectively testable instructions, such as including a specified number of keywords or meeting a length requirement. It contains approximately 500 prompts spanning 25 verifiable instruction types. Follow Bench evaluates content, situation, style, format, and example-based constraints at progressively more difficult levels. These resources show that instruction following should be evaluated separately from response fluency and factual accuracy. arxiv+1

This systematic review examines intent hallucination as an instruction-alignment problem in LLMs. It focuses on four questions:

1. What forms of instruction-following failure are reported in current research?
2. Which benchmarks and metrics are used to detect and evaluate these failures?
3. What technical and prompting-based strategies have been proposed to reduce them?
4. What research gaps remain for multi-constraint, multilingual, conflicting, and multi-turn instructions?

The objective is to provide a structured taxonomy that can support the development of LLM systems that are more reliable, transparent, and aligned with human intent.

II. LITERATURE REVIEW

Instruction-following and intent alignment

Instruction following refers to an LLM's ability to transform a user request into an output that respects stated requirements. These requirements may concern the requested task, factual scope, tone, language, structure, output length, safety boundaries, formatting, or examples. In practice, instruction following involves both understanding and execution: the model must first infer what the user requires and then generate an answer that meets those requirements.

The term intent hallucination is not yet standardized across all LLM research. Related concepts appear under labels such as instruction-following failure, constraint violation, prompt non-compliance, goal misalignment, and requirement omission. The common idea is that the model generates an answer that appears responsive but does not fully satisfy the user's actual request.

A key distinction is therefore necessary. Factual hallucination concerns whether an answer is true; intent hallucination concerns whether it does what was asked.

These two errors can occur independently. A model may provide incorrect information while following every format instruction, or it may provide correct information while violating multiple user constraints. Reliable AI systems need to reduce both forms of failure.

Types of instruction-following failure

The literature suggests that intent hallucination can be categorized into several recurring forms:

Failure type	Description	Example
Omission failure	The model ignores one or more explicit requirements	User asks for advantages and disadvantages; model provides only advantages
Misinterpretation	The model assigns an incorrect meaning to an instruction	"Write for children" is interpreted as using childish rather than simple language
Format violation	The output does not follow required structure or syntax	User requests valid JSON, but model adds explanatory text
Constraint conflict failure	The model follows one instruction while silently violating another	It produces a short answer but leaves out mandatory citations
Priority failure	The model follows a lower-priority instruction over a more important one	It prioritizes style over an explicit safety or system requirement
Multi-turn forgetting	Constraints introduced earlier	A previously selected language or format changes in later turns
Over-compliance	The model follows an Implied pattern too rigidly and adds restrictions the us er did not request	It refuses a harmless request because it incorrectly assumes a prohibited intent
Dynamic-update failure	The model does not properly incorporate revised user instructions	User changes "use a table" to "use bullets," but the model continues using a table

Follow Bench was designed to measure several of these difficulties by organizing constraints into content, situation, style, format, and example categories. It increases task complexity by adding constraints incrementally, allowing researchers to observe how compliance changes as instructions become more demanding.

Evaluation benchmarks

IFEval is one of the most influential benchmarks for instruction-following research. It focuses on instructions that can be verified automatically, such as using a specified number of sections, avoiding certain punctuation, including exact phrases, or meeting a minimum word count. This design reduces dependence on subjective human ratings or LLM-based judges.

However, IFEval mainly evaluates single-turn, English-language prompts with objectively testable constraints. Therefore, it cannot fully represent realistic human–AI interaction, where instructions may be ambiguous, contradictory, culturally dependent, or revised over time.

FollowBench addresses some of these limitations by measuring fine-grained compliance across five major constraint groups. It uses both strict and partial satisfaction measures, helping distinguish complete compliance from partial adherence.

Multi-IF extends instruction-following evaluation into multi-turn and multilingual settings. The benchmark contains 4,501 multilingual conversations, each structured across three turns, and evaluates whether models maintain compliance as instructions evolve during interaction. M-IFEval similarly extends evaluation beyond English by including French, Japanese, and Spanish prompts. Reported results indicate that performance can differ substantially across languages and instruction categories.

These benchmarks demonstrate that a model's strong performance on simple rule-based prompts does not guarantee reliable instruction following in practical scenarios.

Detection and mitigation methods

Detection methods generally fall into three groups:

- Rule-based verification:

Programmatic checks evaluate directly observable requirements, such as word count, output format, keyword presence, or prohibited terms.

- LLM-as-a-judge evaluation:

A separate model evaluates whether the generated response satisfies the prompt. This approach can assess semantic constraints but may introduce judge bias or inconsistency.

- Human evaluation:

Human annotators assess whether the model understood and fulfilled the request. Human evaluation captures nuance but is expensive and may vary across reviewers.

A promising technical direction is to represent an instruction as a set of explicit constraints and evaluate the final answer as a constraint-satisfaction problem. Neuro-symbolic verification research proposes formalizing instruction compliance in this way, enabling more transparent detection of unmet requirements.

Mitigation strategies include instruction tuning, prompt decomposition, self-review, constrained decoding, retrieval of task-relevant rules, and external verification. Self-reflection approaches ask a model to inspect its draft for violations before producing a final answer. Selective reasoning approaches attempt to determine when extended reasoning helps and when it instead reduces compliance. Recent research reports that chain-of-thought prompting can sometimes lower adherence to user constraints, while self-reflection, in-context correction examples, and selective reasoning may partially recover performance.

III. METHODOLOGY

Review design

This study uses a systematic literature review methodology to identify, classify, and synthesize research related to intent hallucination and instruction following in LLMs. The review follows four stages: search, screening, data extraction, and thematic synthesis.

The review treats "intent hallucination" as an umbrella term that includes instruction omission, instruction misinterpretation, constraint violation, multi-turn instruction loss, and failure to resolve conflicting requirements.

Search strategy

Relevant studies should be collected from academic databases and preprint repositories, including:

- IEEE Xplore
- ACM Digital Library
- Scopus
- Web of Science
- Google Scholar
- arXiv
- ACL Anthology
- NeurIPS and ICLR proceedings

Example search strings include:

- "Large language model instruction following"
- "LLM instruction compliance"
- "LLM constraint following"
- "Intent hallucination in language models"
- "Instruction following benchmark LLM"
- "Prompt adherence evaluation"
- "multi-turn instruction following"
- "LLM mitigation instruction failure"

The search should include studies published from 2020 onward because the rapid growth of instruction-tuned generative models occurred mainly after this period.

Inclusion and exclusion criteria

Studies should be included when they:

- Focus on LLM instruction following, prompt compliance, constraint satisfaction, or user-intent alignment.
- Introduce or evaluate a benchmark, detection method, metric, or mitigation method.
- Present empirical findings, a reproducible methodology, or a well-defined conceptual framework.
- Are peer-reviewed papers, credible conference publications, technical reports, or influential preprints.

Studies should be excluded when they:

- Address only factual hallucination without discussing instruction adherence.
- Focus exclusively on traditional dialogue systems without modern LLM relevance.
- Lack sufficient methodological detail.
- Are duplicates, short abstracts, tutorials, or non-scholarly opinion pieces.

Data extraction

For each selected study, the following information should be recorded:

Data field	Description
Bibliographic information	Author, year, venue, and publication type
Research objective	Main problem addressed by the study
Definition of failure	How instruction-follow in errors is described
Task setting	Single-turn, multi-turn, multilingual, tool use, or agent setting
Benchmark or dataset	Evaluation resource used or introduced
Evaluation method	Rule-based, human, LLM judge, or hybrid approach
Metrics	Accuracy, satisfaction rat, violation rate, or human preference
Mitigation technique	Prompting, fine-tuning, verification, decoding, or se lf-refinement
Key findings	Main reported outcomes
Limitations	Weaknesses identified by the authors reviewer

Analysis method

The extracted studies should be analyzed using thematic synthesis. First, each reported failure should be coded according to its type. Second, evaluation methods should be grouped by their degree of automation and semantic coverage. Third, mitigation methods should be categorized according to whether they act before generation, during generation, or after generation.

This process produces a three-part taxonomy:

1. Failure taxonomy: omission, misinterpretation, formatting, prioritization, memory, conflict, and dynamic-update errors.
2. Evaluation taxonomy: deterministic rule-based tests, LLM judges, human evaluation, and hybrid verification.
3. Mitigation taxonomy: prompt-level, training-level, decoding-level, and post-generation verification methods.

IV. RESULTS

Taxonomy of intent hallucination

The reviewed research indicates that instruction-following failure is not one single phenomenon. Instead, it is a collection of related alignment problems. The most common issue is partial compliance: the model fulfills the central task but misses secondary requirements such as output format, length, audience, exclusions, or required examples.

A second major finding is that constraint complexity strongly affects performance. Models generally perform better on simple instructions that can be evaluated with direct rules. Performance becomes less reliable when prompts combine multiple requirements, contain subtle semantic restrictions, or change across multiple conversation turns. FollowBench specifically evaluates this progressive difficulty by incrementally increasing the number of constraints.

A third finding is that multilingual instruction following remains uneven. M-IFEval reports substantial variation across languages and instruction types, showing that high English-language performance should not be assumed to transfer directly to other linguistic settings.

Evaluation methods

Rule-based evaluation is highly useful for objectively verifiable requirements. It is reproducible, inexpensive, and less subjective than human or LLM judging. IFEval demonstrates this approach through automatic checks for requirements such as word count, phrase repetition, and formatting conditions.

However, rule-based methods have limited ability to assess meaning. They can determine whether a response contains three headings, but not necessarily whether the headings are useful, whether the explanation is appropriate for the intended audience, or whether a summary accurately reflects the source material.

LLM-as-a-judge approaches offer broader semantic evaluation, but their reliability depends on the judge model, evaluation prompt, scoring rubric, and possible shared biases between the tested model and the evaluator. Human assessment remains valuable for nuanced cases, especially where requirements are ambiguous or context-sensitive, but it introduces cost and inter-rater variability.

The evidence suggests that hybrid evaluation is the most practical direction. A robust framework can use deterministic checks for measurable constraints, human or LLM evaluation for semantic alignment, and explicit error labels for diagnosing why a response failed.

Mitigation strategies

The literature identifies four broad mitigation categories:

Category	Main approach	Strength
Prompt-level methods	Use clear formatting, task decomposition, examples, and checklists	Easy to deploy without retraining
Training-level methods	Fine-tune models on diverse instruction-following data	Can improve general behavior
Generation-level methods	Apply constrained deco ding or structured output generation	Effective for format compliance
Post-generation methods	Use self-review, extern al validators, and revision loops	Detects errors before delivery

Self-review is especially relevant for complex tasks. The model can be prompted to extract a checklist of requirements, draft a response, verify each requirement, and revise only the sections that violate constraints. Research on efficient self-review frameworks proposes decomposing instructions into task and constraint components before conducting structured revision.

Another emerging direction is conflict-aware instruction processing. Some user prompts contain incompatible requirements, such as asking for both a one-sentence answer and a detailed five-paragraph explanation. Models should not silently choose one requirement.

Instead, they should identify the conflict, ask for clarification when necessary, or state which constraint they are prioritizing. ConInstruct evaluates both conflict detection and conflict resolution, noting that models may recognize contradictions but still fail to explicitly communicate them to users.

Research gaps

Several limitations remain in current research:

- Existing benchmarks often emphasize easily verifiable surface constraints instead of deeper user goals.
- Single-turn English evaluations do not adequately represent real conversational use.
- Few datasets include ambiguous, conflicting, emotional, culturally specific, or evolving user instructions.
- There is no universally accepted definition or measurement standard for intent hallucination.
- Benchmark scores may overestimate real-world reliability because models can learn benchmark patterns.
- More research is needed on long-context, tool-using, agentic, and multimodal systems.
- Evaluation should measure not only whether a constraint was followed, but whether compliance damaged answer quality or usefulness.

V. CONCLUSION

Intent hallucination is an important but underdeveloped area of LLM reliability research. It describes situations in which a model produces an answer that may be fluent or factually accurate but does not fully satisfy the user's stated requirements. These failures include omitted constraints, incorrect interpretations, formatting mistakes, multi-turn memory loss, failure to identify conflicts, and inconsistent handling of updated instructions.

The reviewed literature shows that instruction following can be evaluated through rule-based benchmarks, LLM judges, human assessment, and hybrid frameworks. IFEval provides reproducible tests for objectively verifiable requirements, while FollowBench evaluates fine-grained constraints of increasing complexity. Multi-IF and M-IFEval demonstrate the importance of testing multi-turn and multilingual instruction following rather than relying only on single-turn English benchmarks.

Current mitigation techniques include improved prompt design, instruction tuning, structured generation, self-reflection, selective reasoning, external validation, and constraint-satisfaction approaches. No single method completely solves the problem. The strongest practical approach is likely a layered system that explicitly extracts requirements, generates an answer, checks compliance using appropriate validators, detects conflicts, and revises the output before presenting it to the user.

Future research should develop standardized definitions of intent hallucination, multilingual and multi-turn benchmarks, conflict-aware datasets, and evaluation methods that jointly measure factual accuracy, instruction adherence, safety, and response quality. Improving these capabilities will be essential for creating LLM systems that are dependable in education, healthcare, business, public services, software engineering, and other high-impact applications.

REFERENCES

[1] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," arXiv preprint arXiv:2311.07911, 2023.

[2] Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu, and W. Wang, "FollowBench: A multi-level fine-grained constraints following benchmark for large language models," arXiv preprint arXiv:2310.20410, 2023.

[3] H. He et al., "Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following," arXiv preprint arXiv:2410.15553, 2024.

[4] M-IFEval Authors, "M-IFEval: Multilingual instruction-following evaluation," arXiv preprint arXiv:2502.04688, 2025.

[5] Z. Qin et al., "InfoBench: Evaluating instruction following and information-seeking behavior in large language models," in Proc. NLP Conf., 2024.

[6] Y. Wen et al., "ComplexBench: Evaluating complex instruction following in large language models," arXiv preprint, 2024.

[7] X. Han, "MultiTurnInstruct: Evaluating instruction following in multi-turn dialogue," arXiv preprint, 2025.

[8] C. White et al., "LiveBench: A challenging, contamination-free LLM benchmark," arXiv preprint, 2024.

[9] Y. Zhang et al., “Strengthening LLMs’ complex instruction following with constraint-based feedback,” arXiv preprint, 2025.
[10] X. Li et al., “Neuro-symbolic verification on instruction following of LLMs,” arXiv preprint, 2026.
[11] H. Zhang et al., “When thinking fails: The pitfalls of reasoning for instruction-following in LLMs,” OpenReview, 2026.
[12] Authors, “ConInstruct: Evaluating large language models on conflict detection and resolution in instructions,” arXiv preprint, 2026.
[13] Authors, “Efficient self-review framework for enhancing instruction following,” *OpenReview*, 2026.

Follow-ups
Compare IFEval, FollowBench, Multi-IF, and InfoBench benchmarks constraint types, multi-turn scope, and scoring metrics
Computer

Instruction-following mitigation techniques compared: in-context learning vs self-reflection vs neuro-symbolic CSP verification
Computer

What are the main benchmarks used to evaluate intent hallucination
How does intent hallucination differ from factual hallucination in LLMs
What mitigation techniques are most effective against constraint violations