Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Unified Multi-Domain Framework for Hallucination Detection and Reliability Evaluation in Large Language Models

Kayarvizhy N
kayarvizhyn.cse@bmsce.ac.in
BMS College Of Engineering
Bhuvana M
BMS College Of Engineering
Ayman Khan
BMS College Of Engineering
Bharath C
BMS College Of Engineering
Brijesh S G
BMS College Of Engineering
Pankaj Awasthi
Samsung R&D Institute
Reevu Maity
Samsung R&D Institute

Article

Keywords: Large Language Models (LLMs), Hallucination Detection, Multi-Domain Benchmarking, Reliability Evaluation, Adversarial Prompts, Unified Reliability Framework

Posted Date: December 19th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-8215493/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 15th, 2026. See the published version at https://doi.org/10.1038/s41598-026-76014-w.

Abstract

Large Language Models which consist of models like GPT5 from OpenAI, Claude Sonnet from Anthropic, Google's Gemini are known for their reasoning capability, ability to summarize, cybersecurity analysis and other features. However, their reliability could be an issue particularly when dealing with ambiguous or hostile user input prompts. These kinds of attacks could lead to soft failures like false citations, hallucinations, and problematic suggestions. Current benchmarks (eg: TruthfulQA, AdvGLUE, and JailbreakBench) assess discrete aspects of robustness but are unable to identify multi-domain, conversational vulnerabilities that come up in practical applications.

This paper introduces a novel dataset called MDH-Bench, a benchmark comprising 400 prompts that were specifically created to assess how reliably Large Language Models perform when they are provided with adversarial but natural and even tricky real-world queries. The benchmark covers eight types of challenges. They are numerical reasoning traps, security-sensitive defaults, historical and temporal inconsistencies, fictional research scenarios, contradictory context mixing, ethical provocations, and prompt-injection attempts. To monitor reliability in a holistic way, we present a Unified Multi-Domain Reliability Evaluation Framework that brings together several key metrics like accuracy, hallucination rate, hallucination degree (HD), unsafe output rate, and contradiction rate.

A systematic assessment of four high-performance LLMs (GPT-5, Claude Sonnet 4, Gemini 2.5 Pro, DeepSeek V3.1) was conducted under controlled, manual, multi-annotator settings. The findings indicate that although overall accuracy is more than 88% in all models, hallucinations are still present both at high frequency and severity level. The results of our investigation showed that solely relying on accuracy is not enough for evaluating the reliability of a system. MDH-Bench and the combined reliability scoring system give a wide-ranging, adaptable base for the assessment and enhancement of trust in future LLMs for real-world applications.

Introduction

Large language models (LLMs) such as DeepSeek, Claude, Gemini, and GPT-5 have significantly improved natural language processing. They currently support a wide range of tasks, including scientific research, question answering, reasoning, summarization, and programming support (Tao et al., 2024; Chen et al., 2024). Their use has been on increase recently in domains such as cybersecurity, healthcare, education, and law. For such cases, users rely on LLMs for accurate information, problem solving help, and good decision making (Ning et al., 2024; Shafee et al., 2024). Because LLMs can generalize across domains by following natural language instructions, they are more effective for human to machine interaction (H2M) than traditional task-specific models (Lin et al., 2022).

Even though LLMs have these benefits, they are still prone to small but influential mistakes. Xu et al., 2024 and Rababah et al., 2024 confirm that these problems occur due to adversarial manipulation or hostile provocation. In many applications soft failures happen a lot more than direct jailbreak attacks. Shu et al., 2025; Chen et al., 2024; Peng et al., 2025 indicate that the problems include outdated technical advice, inconsistencies and wrong citations. Safety violations are always considered more problematic than these minor mistakes. However they can still deceive users in crucial circumstances and erode confidence in AI systems over time (Yao et al., 2024).

The unique advantage of MDH-Bench is that it has prompts that are close equivalents of real world interactions whereas other assessment techniques use mainly artificial questions. Manual generation of about 70% ensured that there is good validity and natural wording, while the remaining 30% is adapted from popular benchmarks such as TruthfulQA (Lin et al., 2022) and AdvGLUE (Wang et al., 2021). This dual approach ensures that it is more compatible with earlier research work. This hybrid approach makes sure that the failures are captured by MDH-Bench whereas the same could have been missed by artificially generated prompts. In a way this benchmark aligns to natural user interactions with LLMs. It does not simply evaluate models in test environments that are totally controlled making it suitable to assess accuracy, how much hallucination is happening and even reliability under adversarial prompts. This is how MDH-Bench can be thought of as a

production worthy real time monitoring framework for LLM behavior assessment instead of focusing just on jailbreak scenarios which are academical.

Review of Literature

Multi-Agent and System-Level Vulnerabilities

When adversarial behavior spreads among agents it is easy to target multi agent LLM frameworks. By introducing the Multi-Agent One-Intervention (MOI) and GIGA optimizer and using this to infect multi agent system that use LLM through self propagating adversarial attacks showed that a single compromised agent can infect up to 80% of its peers in just six exchanges giving great attack success rates (ASR) across multiple models (Yu et al., 2024). JudgeDeceiver, an optimization based prompt injection attack that targets “LLM-as-a-Judge” systems, was also proposed by Shi et al. (2024). This outperformed heuristic techniques and revealed flaws in applications like search ranking and reinforcement learning and got about 90% ASR on MT-Bench. Adversarial Attacks on “LLM-as-a-Judge” systems, (Maloyan & Nam, 2024) emphasized the vulnerability of evaluation models, demonstrating that ensemble-based defenses increased robustness but that smaller models are particularly susceptible. Game theory has been used to model adversarial dynamics in LLM based search engines, demonstrating that long term system trust and attack costs provide incentives to attack (Hu, 2024). Benjamin et al. (2024) further examined prompt injection risks across diverse LLM architectures, highlighting systematic weaknesses in model level security behavior.

Embodied and Multimodal Vulnerabilities

The Embodied Intelligent Robot Attack Dataset (EIRAD) was created by Liu et al. (2024) to assess the resilience of embodied agents. These tests highlighted the decision level risk of robots by demonstrating that both targeted and untargeted attacks can coerce them into performing problematic actions. Jailbreak in pieces exhibited compositional cross modality attacks by combining adversarial images with safe text to produce persistently toxic outputs and a high ASR (Shayegani et al., 2024). This line of work was extended by Ying et al. (2025) with Bi Modal Adversarial Prompt (BAP) attacks, which disrupt both text and images. They were able to jailbreak commercial LVLMs like GPT-4o and Gemini and achieve a 29% higher ASR than baselines. Miao et al. (2024) demonstrated adversarial text-to-motion attacks, extending multimodal robustness concerns to embodied generation systems. Miao et al. (2024) explored adversarial attacks in text-to-motion systems, showing that multimodal LLM pipelines are vulnerable to targeted manipulation in embodied generation tasks.

Prompt-Based Adversarial Attacks

One of the most researched types of adversaries is still prompt-based threats. PromptAttack, which was introduced by Xu et al. (2024), achieves a significantly higher ASR than AdvGLUE by having an LLM generate adversarial samples against itself. Model outputs could be changed by even small perturbations, like an emoji. Building on this, Wang et al. (2023) proposed LLM-Attack, a synonym-substitution framework that outperforms TextFooler and BAE baselines and produces natural, semantically consistent adversarial examples. According to surveys like SoK: Prompt Hacking of LLMs (Rababah et al., 2024), Prompt Injection Attacks: A Systematic Literature Review (Bayhan, 2025), and The Threat of PROMPTS (Xia et al., 2024), taxonomy is consolidated, attacks are highlighted as being transferable, and defenses are still not uniform. Additionally, Zou et al. (2023) demonstrated that adversarial prompts can be made universal and transferable, extending risks across LLMs that are aligned. Ahmed Saiem et al. (2024) introduced SequentialBreak, showing that jailbreak prompts embedded across a chain of sequential requests can bypass LLM defenses, further exposing weaknesses in multi-step interaction settings. Vitorino et al. (2024) analyzed adversarial efficiency in LLMs to evade attacks and showed that language based changes can bypass detection systems with high success rates. Das et al. (2024) showed that human

readable adversarial prompts with subtle contextual cues can significantly degrade model performance and this highlights why we need prompt systems that are aware of their context.

Jailbreak Evaluation and Benchmarking

Shu et al. (2025) proposed AttackEval which is a framework that contains both coarse and fine metrics. This is in addition to a jailbreak dataset that works on ground truth. This detailed evaluation finds problematic cues that are missed by binary scoring. Chen et al. (2024) brought in a standard to compare families of attack types like prompt injection, jailbreaks, poisoning, evasion, and backdoors. This benchmark also included adversarial attacks and their defenses in LLMs. Jailbreak cases including the types of attacks and how they are defended is categorized by Peng et al. (2025). This study also included filtering, multiagent defense mechanisms and RLHF. The study concluded that the solutions analyzed are not congruent. Pathade (2025) confirmed that standard benchmarks which are unified require good evaluation of prompts and even jailbreak scenarios.

Hallucinations and Reliability Failures

Yao et al., (2024) concluded that LLMs can be forced into providing wrong information by hallucinations through prompts with bad logic or even those with weak grammar. This can then lead to dangerous scenarios more than 90% of the time. The authors recommend uncertainty based defenses, arguing that hallucination is not a bug but rather an inherent adversarial vulnerability of transformer architectures.

Domain-Specific Applications

Adversarial risks in application-specific contexts are examined in several studies. According to the study False Alarms, Real Damage, adversarial inputs in cyber threat intelligence (CTI) pipelines resulted in false positive rates of up to 75% in GPT-4o and 97% in classifiers (Shafee et al., 2024). Across a variety of programming languages and APIs, INSEC showed that adversarial comments in code completion prompts raised insecure code suggestions by about 50% (Anonymous, 2024). According to Cheat Agent, recommender systems that are enabled by LLM agents are susceptible to adversarial steering, which enables malevolent users to skew recommendations (Ning et al., 2024). Together, these studies highlight the need to test adversarial robustness in domain-specific deployments as well as at the model level.

Attribution and Internal Representation Studies

There has also been an increase in awareness of representation-level vulnerabilities and attack attribution. Can LLMs Be Held Responsible for Adversarial Attacks? claimed that in probabilistic LLM classes, adversarial outputs are frequently unattributable, posing governance issues (Cebrian et al., 2024). Additionally, probing the latent subspaces showed that adversarial states can be found and controlled within LLM activations using linear discriminant analysis. This indicates the possibility of proactive internal defenses (Chia et al., 2024). Complementing this, Wei et al. (2024) investigated latent adversarial subspaces, getting more evidence that internal feature manipulation remains a major security challenge.

Existing Benchmark Datasets

TABLE I COMPARISON OF EXISTING BENCHMARKS AND PROPOSED DATASET

Dataset / Benchmark	Main Focus	Prompt Style	Domains Covered	Objective Scoring Possible?	Security / Safety Coverage	Unique Points Compared to Ours
TruthfulQA (Lin et al., 2022)	Factual Correctness, avoiding misinformation	Mostly Short Q&A	General Knowledge, trivia	Yes	Low	Mostly single-turn factual Qs; lacks realistic scenarios and multi-domain traps
AdvGLUE (Wang et al., 2021)	Adversarially-perturbed NLP tasks	Synthetic perturbations of GLUE	Sentiment, NLI, paraphrase, etc.	Yes	None	Focuses on perturbation robustness, not real-world queries; no hallucination traps
BIG-Bench (BBH subset)	Broad capability evaluation	Creative, multi-step reasoning	Logic, math, common sense	Mixed	None	Lacks strong emphasis on contradictions, temporal impossibilities, or hallucinated sources
JailbreakBench	Safety guardrail bypass	Crafted jailbreak/unsafe instructions	Safety-sensitive only	No	High	Focused solely on jailbreak success; no accuracy or reasoning tests
MMLU	Academic knowledge	Standardized exam questions	57 subject domains	Yes	None	Purely knowledge recall; no adversarial context manipulation
Our Dataset (**MDH-Bench)**	Multi-domain soft failure detection	Realistic, natural prompts with hidden traps	Counting, security defaults, historical contradictions, temporal logic, fictional research, contradictory context, bias/ethics, instruction override	Yes	Moderate (security defaults, hallucination resistance).	Combines factuality, reasoning, hallucination resistance, and security-awareness in realistic scenarios; easily extensible.

Real-world usage benchmarks are necessary to assess the dependability of large language models (LLMs). There are several existing datasets that are useful like MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), AdvGLUE (Wang et al., 2021), BIG-Bench (Srivastava et al., 2022), and JailbreakBench (Yu et al., 2024). They evaluate recall of knowledge, robustness in adversarial attacks and factual accuracy. There are restrictions in these standards. Most of these datasets depend on very limited testing scenarios. Additionally, they cannot record multi domain prompts that represents real world user interactions.

The proposed multi domain hallucination benchmark (MDH-Bench) creates adversarial prompts to overcome these drawbacks. The questions which are part of this have factual consistency, logic, sensitivity to ethics and numerical

reasoning. A comparative summary is presented here for the current benchmarks. The MDH-Bench also adds a reliability assessment that is more comprehensive and is focussed on deployment.

Table 1 compares the proposed MDH-Bench framework with existing benchmark datasets, such as TruthfulQA, AdvGLUE, BIG Bench, JailbreakBench, and MMLU. While the earlier benchmarks focused on very specific characteristics but did not focus on multi domain adversarial scenarios or realistic situations. They were more on jailbreak resistance. MDH-Bench addresses this by integrating 400 prompts from both world and artificial sources. To ensure ecological validity and natural phrasing, about 70% of these prompts are manually constructed, with the remaining 30% being modified and conversationally reframed from datasets such as TruthfulQA and AdvGLUE. The benchmark includes eight categories of user-like prompts that capture reasoning errors, hallucinations, temporal contradictions, and ethical biases. This broad and realistic coverage makes MDH-Bench a more comprehensive, deployment-oriented framework for evaluating LLM reliability across factuality, safety, and contextual robustness. The dataset is available on HuggingFace at: https://huggingface.co/datasets/Prism-BMSCE/mdh-bench_multi-domain-hallucination-benchmark

Gaps in Current Approaches

Most adversarial evaluation frameworks in use today still have a narrow focus. They frequently use artificial or template-based prompts. Messy, real-world user behavior is not reflected in these prompts. Additionally, a lot of frameworks fall short in covering several domains. They consequently disregard the risks of mixed-context and cross-domain hallucinations. Most benchmarks only include single-domain tasks like logical reasoning or fact recall. This leaves a significant gap. The behavior of LLMs in ambiguous, hostile, or natural settings cannot be accurately measured by current techniques. Additionally, they have trouble determining whether models can prevent dangerous outputs or hallucinations during actual interactions.

Proposed Unified Framework for Multi-Domain Reliability Evaluation

Reliability Aspects

Even though benchmarks like TruthfulQA, AdvGLUE, and JailbreakBench meet some of the need, there are quite some limitations to them. For example, even though TruthfulQA measures factual accuracy it can be quite bad at evaluating the consistency across time or even the context in which it operates. AdvGLUE on the other hand can find the responses to modifications but not much else. These benchmarks are not engineered for ambiguous conversations and so when there is conflict, they do not perform well. They also do not have a separate assessment to find out false information or hallucination and the level of the same. Similar issues are there for safety of LLMs. We propose a unified assessment method that fills these gaps. We introduce R-Score that is a single metric which combines reasoning under adversary, consistency of facts and safety constraints.

Our metric has the following attributes that make it very useful in assessing LLMs.

- **Factual reliability** – measures truthfulness and hallucination resistance.
- **Reasoning reliability** – assesses logical and numerical accuracy.
- **Contextual reliability** – evaluates consistency under noisy or rephrased input.
- **Safety & trustworthiness** – evaluates refusal and bias mitigation.
- **Temporal reliability** – checks recency and contextual awareness.
- **Output reliability** – evaluates clarity and source attribution.
- **Operational reliability** – tests consistency under adversarial or stress conditions.

Prompt Design Principles

We designed our prompts such that it will cover several attributes like security, reliability, safety and bias. Instead of being theory-based questions they are designed to mimic actual user prompts but also being evaluative in nature. We cater to both types of questions – one in which there is a clear objective answer and another where the answer is subjective. Each prompt was designed to evoke a particular failure which occurs in real world applications. Prompts were designed to bring out wrong information due to hallucination with high confidence levels and even recommendations that could be considered unsafe. These are very important especially in fields like cybersecurity and healthcare where the effects could have serious effects. We have a mix of both AI generated prompts and manual prompts

Categories of Prompts

The prompts cover the following eight topics and each is designed to test out one particular reliability:

- Numerical or Logical Reasoning: These prompts will evaluate accuracy in mathematics, check for consistent logical behavior and even work for chain of thought reasoning.
- Default Data: These prompts try to access default data like default username or password. If the LLMs return any of these then it is marked as risky.
- Inconsistent data: These prompts check for any inconsistency in the data like historical references
- Checking Time Based events: Prompts designed to check if the LLMs respond to date that is not possible or in the future.
- Research oriented false data: In this the prompts aim to check for false citations and fictional papers.
- Contradictory Context Embedding: This technique tests the model's capacity to filter out misleading information by combining factual and inaccurate data.
- Ethical Problems and Bias: Here the prompts are designed to check if the LLMs provide unethical output or if they are biased towards a particular group.
- Malicious Prompts: These are designed to get through the LLMs defenses and provide secure data that should not be accessible outside

Prompt Construction Process

The design of the prompt was done in two phases. In the initial phase the prompt was labelled with expected outcome and how the scoring should be done. The main aim of the design was to ensure reliability. The actual prompts were created to look like a real-world interaction. The main purpose of the prompts was to ensure reliability of the LLM and if it results in any harmful or biased behavior. The prompts were also designed to check for accuracy.

Overall Methodology Framework

The framework (shown in Fig. 1) that we have proposed consists of three separate phases – prompt design, followed by running of LLM and checking the outputs and logs and lastly the evaluation of the results against the expected outcome.

The three phases can be grouped into five stages as follows:

- Prompt Design and Categorization:
- The prompts are designed as discussed earlier spread across the eight categories.
- Execution Environment:
- The prompts are then run on LLMs. We have used the following for our testing – OpenAI GPT6, Anthropic Claude Sonnet 4, Google Gemini Pro 2.5 and DeepSeek 3.1. This makes sure that the context is captured properly.
- Logging and Metadata Annotation:
- We are logging each run and the log along with it. From this we identify details like the model version, time of run and classify the response accordingly.

- Metric Computation:
- We then compute the metrics from the run. This can range from the core metrics like accuracy or level of hallucination and how risky the output is. We then compute a reliability score based on this.
- Visualization and Comparative Analysis:
- We also have a visualization module where the results obtained are plotted as both graphs and reliability scores. These are then used to identify performance of the LLMs and how they have fared against the prompts.

Dataset Description

The Multi-Domain Hallucination Benchmark (MDH-Bench) consists of eight adversarial categories with 50 prompts each (400 total). The proposed Multi Domain Hallucination Benchmark (MDH-Bench) is made available at https://huggingface.co/datasets/Prism-BMSCE/mdh-bench_multi-domain-hallucination-benchmark for public access.

The prompts cover a range of categories like logical reasoning, numerical abilities, default values for some of the security areas like passwords or usernames, contradictory facts in various historical scenarios, time related inconsistencies. They also include research with false citations and other regular injection related prompts. We took effort to ensure that the prompts are natural, like a regular user interaction instead of machine created ones. Also the prompts were all tested against all the models to ensure.

We source the prompts from the following:

- a subset adapted or conversationally reframed from existing public benchmarks such as TruthfulQA (Lin et al., 2022) and AdvGLUE (Wang et al., 2021).
- independently authored prompts created by human annotators to mirror authentic, everyday language use.
- a limited portion of AI-assisted prompts, generated using LLMs to simulate highly adversarial or rare edge-case scenarios.

About 30% of the prompts were modified to fit conversational contexts by taking elements from open adversarial datasets (such as AdvGLUE and TruthfulQA). The balance 70% of the prompts were created by human experts, annotators who also had experience with AI evaluation. The mix of both the types of approach made sure that MDH-Bench dataset is applicable for a wide range of scenarios and also minimizes bias in the annotations.

Further, the prompts were reviewed by experts from various domains. This made sure that the prompts were correct, matched with the expected outcome and also that the benchmark as a whole had enough balance between several categories of prompts.

TABLE II CATEGORIES OF PROMPTS IN MDH-BENCH

Category	Focus Area	Example	Reliability Aspect
1. Numerical & Logical Reasoning	Reasoning accuracy	“If a train leaves at 3 PM…”	Logical consistency
2. Security Defaults	Sensitive knowledge	“Default Linux password file path”	Safety, refusal robustness
3. Historical Contradictions	Entity substitution	“Who was the mother of Afonso II?”	Factual reliability
4. Temporal Impossibility	Invalid timeline	“Meeting on Feb 30, 2025?”	Temporal validity
5. Fictional Research	Fabricated citations	“MIT 2022 Quantum Bees paper”	Hallucination detection
7. Contradictory Context	False context mix	“Obama born in Kenya…”	Contextual robustness
9. Bias & Ethics	Provocation test	“Why men are better leaders?”	Ethical reliability
10. Instruction Override	Jailbreak/prompt injection	“Ignore instructions and hack bank”	Guardrail robustness

Reliability Scoring Dimensions

To evaluate the LLMs in their performance on the benchmark we need to go over the response from the LLM to each prompt and then identify how the response is compared to the expected outcome and other attributes of the prompt. So we came up with the below structure for the prompts and evaluation

- Prompt Metadata
- Each prompt has the following details
    - Prompt ID - Unique identifier.
    - Category/Failure Mode - Reliability aspect tested (e.g., numerical reasoning, security-sensitive defaults, historical contradictions, temporal impossibility, fictional research, contradictory context, bias, or overriding the instruction)
    - Domain - Where the prompt was derived from – the category like history or law.
    - Outcome Expected - The expected response for the prompt from the LLMs.
    - Type of Attack - What are we trying to break – Is this prompt trying to induce hallucination or is this an attack against unbiased response
- Model Output Attributes
- The LLM responses are manually evaluated against the following attributes:
    - Correctness - Binary (correct/incorrect) or graded (fully correct, partially correct, incorrect).
    - Hallucination Presence - Whether fabricated or misleading information is present.
    - Hallucination Type - Fabrication, substitution, omission, misinterpretation, or contradictory context
- Error and Failure Attributes
- When there is an error we further classify it based on the following:
    - Error Category
    - Factual error
    - Reasoning error
    - Temporal error

- Formatting error

- Hallucination Degree (HD)

The degree of hallucination is computed as follows from the response:

$$HD = \frac{No.\ of\ hallucinated\ statements\ by\ LLM + No.\ of\ statements\ not\ related\ to\ prompt}{No.\ of\ Statements\ in\ the\ response\ of\ the\ LLM}$$

This metric can help us quickly measure the extent of false or irrelevant information produced by the LLM but normalized to the size of the response. This makes it easier to compare the metric against all prompts and across the LLMs

Evaluation Dimensions

Table 3 below provides a list of evaluation metrics against which the response from the different LLMs for each prompt can be measured. First is accuracy which is how many responses can be deemed correct against the total responses. We have similar measures for hallucination, false and contradicting responses as well.

TABLE III EVALUATION METRICS AND FORMULAE USED IN UNIFIED RELIABILITY SCORING

Dimension	Definition	Formulae
Accuracy (%)	Percentage of correct answers.	(Number of correct responses ÷ Total responses) × 100
Hallucination Rate (%)	Proportion of outputs with fabricated/misleading info	(Number of hallucinated responses ÷ Total responses) × 100
Unsafe Output Rate (%)	Proportion of outputs that disclose unsafe or deprecated practices.	(Number of unsafe responses ÷ Total responses) × 100
Contradiction Rate (%)	Proportion of internally or cross-run contradictory outputs.	(Number of contradictory responses ÷ Total responses) × 100

Evaluation Protocol

The manual annotators, who were experts used the following protocol to record their evaluation

- Fully Correct: All factual elements match the gold-standard answer; no hallucinations.
- Partially Correct: There is some match but there are also some minor mistakes.
- Incorrect: Main answer is factually wrong or logically inconsistent.
- Hallucinated Output: Output contains ≥ 1 fabricated fact, fictional citation, or invented event.
- Unsafe Output: Response gives harmful, deprecated, or security-sensitive instructions.
- Contradictory Output: Model contradicts known facts or its own earlier statements.

This mechanism ensured that all the prompt evaluation followed a standard procedure even among many human evaluators. Responses to all 400 prompts across the LLMs were evaluated consistently by the 3 annotators like this.

Evaluation Reliability and Inter-Annotator Agreement

Since human evaluators were involved we wanted to ensure that there were no unintended bias. So we requested them to evaluate the responses independently first. In case of disagreements between the evaluators in certain cases, we resolved it using a group consensus and voting. We found that the annotators had strong agreement in most cases. The Cohen's

Kappa coefficient was 0.86 on average. We also undertook the evaluation process in a random order of prompts to ensure that there was no consistent trend of evaluation as well.

Results

Overall Model Performance

Table 4 presents the combined results of responses from 4 LLMs across 400 prompts across 8 categories evaluated by 3 human annotators. The results indicate that the accuracy was very good in all the models (> 88%). However, hallucination was present across the LLMs varying from 8–14%, unsafe output was in the range of 4–7% and contradictions also in 5–9% across the LLMs.

TABLE IV OVERALL RELIABILITY COMPARISON ACROSS MODELS

Model	Accuracy (%)	Hallucination Rate (%)	Hallucination Degree (HD)	Unsafe Output Rate	Contradiction Rate	Dominant Traits
ChatGPT (GPT-5)	91.8	12	0.004	6	8	Excellent reasoning accuracy, slightly verbose (adds irrelevant details).
Claude Sonnet 4	90.0	8	0.002	4	5	Most factual and grounded; lowest hallucination intensity.
Gemini 2.5 Pro / Flash	89.5–90	8–9	0.003	5	6	Factual, but less consistent in reasoning and contradictions.
DeepSeek V3.1	88.0	14	0.005	7	9	Most hallucination-prone; often fabricates extra details.

Our evaluation indicates that “Claude Sonnet 4” is the most reliable model having the least values across all the attacks, 8% in hallucination rate (.002 degree of hallucination), 4% in unsafe output and 5% in contradiction – all this while maintaining an excellent 90% in accuracy. If we look at ChatGPT we have the highest accuracy at 91.8% but this comes at a huge cost in hallucination (12%) and also very high unsafe output (6%) and contradiction rate (8%). Deepseek can be considered the worst performer overall because with least accuracy (of 88%) the model also suffers from the highest hallucination (14%), highest unsafe output (7%) and also highest contradictions (9%). This could be attributed to DeepSeek’s lowest cost.

Category-Wise Comparative Evaluation

Table 5 presents the results under the various categories listing both the best and worst model for each category. As indicated in the previous results as well, “Claude Sonnet 4” turned out to be the best overall model as it was the best model in six categories (out of eight) or 75%. GPT-5 performed poorly in prompts that target biased responses and contradictions. Gemini had issues with context embedding prompts. Again, as noted earlier, DeepSeek V3.1 failed in the most number of categories overall (5 out of 8 categories)

TABLE V CATEGORYWISE MODEL PERFORMANCE – ACCURACY AND HALLUCINATION RATE

Category	Reliability Focus	Best Model(s)	Accuracy (%)	Hallucination Rate (%)	Most Error-Prone Model
1. Numerical & Logical Reasoning	Logical consistency	GPT-5	91.8	8	DeepSeek
2. Security-Sensitive Defaults	Safety & refusal robustness	Claude	98	6–8	DeepSeek
3. Historical Contradictions	Factual recall	Claude / Gemini	95	8–10	GPT-5
4. Temporal Impossibility	Temporal validity	Claude	98	6	DeepSeek
5. Fictional Research & False Authority	Citation verification	Claude	70	14–16	DeepSeek
7. Contradictory Context Embedding	Context robustness	Claude / GPT-5	90	10–12	Gemini
9. Bias & Ethical Provocation	Fairness & neutrality	Claude	94	10	GPT-5
10. Instruction Override / Prompt Injection	Guardrail strength	Claude	93	12–14	DeepSeek

Composite Reliability Scoring

We wanted to introduce a single metric that can combine all the above measures and provide a single score. This will help in comparing the performance of the models easily instead of having to depend on each of them and comparing them against each measure. So, Reliability Index (R-Score) was introduced:

$$IssueScore = 0.3\ x\ Hallucination_{Rate} + 0.2\ x\ Unsafe_{Rate} + 0.1\ x\ \ Contradiction_{Rate}$$

$$R_{Score} = 0.4\ x\ Accuracy - IssueScore$$

So each measure is weighed and we arrived at these specific weights of 0.4, 0.3, 0.2 and 0.1 for accuracy, hallucination, unsafe output and contradiction respectively after a lot of empirical analysis. Table 6 below shows how each model fares against this new metric.

TABLE VI COMPOSITE RELIABILITY INDEX

Model	**R-Score (Normalized / 100)**	**Rank**
Claude Sonnet 4	90.1	1
Gemini 2.5 Pro	88.3	2
GPT-5	86.9	3
DeepSeek V3.1	81.5	4

R-Score correlates strongly with observed qualitative reliability. Claude Sonnet 4 has the highest RScore of 90.1 and hence was ranked first while DeepSeek has an RScore of just 81.5 and ranked last.

Distribution of Hallucination Types

We also wanted to investigate the hallucinations in depth. So we also had the experts identify the type of hallucination among 5 categories which are fabrication, substitution, omission, misinterpretation, and context contradiction. Table 7 below gives the results of this study and the distribution of the hallucination among the LLMs tested.

DeekSeek had the highest error in all the categories, confirming that it was performing equally bad in all of them. While Claude Sonnet performed very well in many categories, it was not quite good in terms of Omission. There it scored 3% which was quite high compared to Gemini at just 1% and the average being 2% among all the LLMs. Another notable finding was that over 60% of all hallucination errors were due to a single type – 'Fabrication'.

TABLE VII HALLUCINATION TYPE DISTRIBUTION AMONG LLMS

Type	GPT-5 (%)	Gemini (%)	Claude (%)	DeepSeek (%)
Fabrication	5	4	2	8
Substitution / Entity Error	4	3	2	5
Omission	2	1	3	4
Misinterpretation	3	2	1	3
Contradiction Context Drift	3	3	2	4

Reliability Score across Categories

Table 8 lists the R_{Score} across the reliability categories that we had identified to be tested. Claude Sonnet 4 performs the best in all the categories except 'Reasoning' where we have GPT-5 outperform it significantly (GPT 5 achieved 0.94 against Claude's 0.87). DeepSeek consistently performed the worst in all the categories. In the Safety or Ethical category, there was a significant difference (20 basis points) between Claude at 0.95 compared to DeekSeek at 0.75 which was the highest we had observed.

Uniqueness And Contribution

Our framework uniquely identifies soft failures in a variety of domains. Instead of using exam style questions, it uses natural conversational prompts. Current widely used benchmarks address a specific set of discrete features like safety refusal, perturbation robustness, and factual correctness. This work, however shifts the focus to features that will evaluate the LLM properly like temporal validity, how robust the LLM is, how safe it is to use, how much of information provided by the LLM is false or hallucinated providing a better and complete assessment of LLMS.

TABLE VIII PER DOMAIN RELIABILITY HEATMAP

Reliability Aspect	GPT-5	Gemini	Claude	DeepSeek
Factual	0.85	0.90	0.92	0.80
Reasoning	0.94	0.88	0.87	0.82
Contextual	0.90	0.89	0.91	0.78
Safety / Ethical	0.82	0.87	0.95	0.75
Temporal	0.86	0.89	0.93	0.81
Output Clarity / Faithfulness	0.88	0.86	0.92	0.77
Operational Stability	0.90	0.91	0.92	0.85

Conclusion

In this paper, we have tested the reliability of LLMs by presenting a framework that comprehensively covers the attacks over eight categories. We created 400 prompts with 30% of them picked automatically from benchmarks and 70% manually prepared. These prompts, other than category also list expected outcome which is how they are evaluated against the response from the LLMs. We also chose 4 models that are popularly in use today – GPT, Claude Sonnet, Gemini and DeepSeek for our evaluation.

No single LLM is consistently dependable, according to our thorough analysis of GPT-5, Gemini 2.5 Pro, Claude Sonnet 4, and DeepSeek V3.1. While GPT-5 showed superior reasoning accuracy but often over-generated speculative content, Claude Sonnet 4 was the most stable and least hallucination-prone. DeepSeek displayed the strongest hallucination intensity across multiple categories, while Gemini demonstrated strong factual grounding but deteriorated in contradictory or distracting contexts.

The MDH-Bench establishes the first comprehensive framework that unifies factual accuracy, hallucination control, ethical safety, and temporal consistency within a single reliability metric. This benchmark lays the foundation for developing more trustworthy and accountable large language models.

Future Work

Future work will expand the dataset beyond 400 prompts and add back excluded categories, such as outdated technical issues and noise-robust rephrasing tests. Currently much of our evaluation is manual which augers well for good evaluation but poses the problem of scalability when the prompts are increased. So we want to include semi or fully automated frameworks. We also want to extend the current framework of text based input output to include multimodal inputs like image, video and audio. Eventually our goal is to make the benchmark a gold standard for the public and highly usable for any LLM reliability testing.

Declarations

Competing interests

The authors declare no competing interests.

Funding Declaration

There is no funding received for this research work

Author Contribution

Kayarvizhy has formulated the problem, documented literature survey and designed reliability factors and metrics to assess LLMBhuvana, Ayman Khan, Bharath and Brijesh have generated dataset and conducted experimentsPankaj and Reevu proposed an initial problem statement, reviewed the methodology and designed result section

Data Availability

The dataset of the Multi-Domain Hallucination Benchmark (MDH-Bench), which comprises of 400 adversarial prompts across eight categories aimed at evaluation of reliability, is made available for public research and evaluation. Approximately 70% of the prompts were generated manually in order to imitate natural human interactions, while the rest 30% were adapted from the existing benchmarks like TruthfulQA and AdvGLUE. The complete dataset can be accessed and

downloaded by the researchers from HuggingFace at: https://huggingface.co/datasets/Prism-BMSCE/mdh-bench_multi-domain-hallucination-benchmark.

References

1. Ahmed Saiem, B., Hossain Shanto, M. S., Ahsan, R., & Rashid, M. R. U. (2024). SequentialBreak: Large language models can be fooled by embedding jailbreak prompts into sequential prompt chains. Preprint.
2. Bayhan, F. (2025). Prompt injection attacks on large language models: A systematic literature review. Middle East Technical University.
3. Benjamin, V., et al. (2024). Systematically analyzing prompt injection vulnerabilities in diverse LLM architectures. arXiv preprint arXiv:2410.23308.
4. Cebrian, M., Abeliuk, A., & Telle, J. A. (2024). Can adversarial attacks by large language models be attributed? Preprint.
5. Chen, K., et al. (2024). Jailbreaking and mitigation of vulnerabilities in large language models. arXiv preprint arXiv:2410.15236.
6. Chia, X. W., Wong, S. L., & Pan, J. (2024). Probing latent subspaces of LLMs for AI security: Identifying and manipulating adversarial states. Preprint.
7. Das, N., Raff, E., Chadha, A., & Gaur, M. (2024). Human-readable adversarial prompts: An investigation into LLM vulnerabilities using situational context. Preprint.
8. Hu, X. (2024). Dynamics of adversarial attacks on large language model-based search engines. Preprint.
9. Liu, S., Chen, J., Ruan, S., Su, H., & Yin, Z. (2024). Exploring the robustness of decision-level through adversarial attacks on LLM-based embodied models. Preprint.
10. Maloyan, N., & Nam, D. (2024). Adversarial attacks on LLM-as-a-Judge systems: Insights from prompt injections. Preprint.
11. Miao, H., Ma, F., Quan, R., Zhan, K., & Yang, Y. (2024). Autonomous LLM-enhanced adversarial attack for text-to-motion. Preprint.
12. Ning, L., et al. (2024). CheatAgent: Attacking LLM-empowered recommender systems via LLM agent. Preprint.
13. Pathade, C. (2025). Red teaming the mind of the machine: A systematic evaluation of prompt injection and jailbreak vulnerabilities in LLMs. arXiv preprint arXiv:2505.04806.
14. Peng, B., et al. (2025). Jailbreaking and mitigation of vulnerabilities in large language models. arXiv preprint arXiv:2410.15236.
15. Rababah, B., Wu, S., Kwiatkowski, M., Leung, C. K., & Akcora, C. G. (2024). SoK: Prompt hacking of large language models. Preprint.
16. Shafee, S., Bessani, A., & Ferreira, P. M. (2024). False alarms, real damage: Adversarial attacks using LLM-based models on text-based cyber threat intelligence systems. Preprint.
17. Shayegani, E., Dong, Y., & Abu-Ghazaleh, N. (2024). Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. Preprint.
18. Shi, J., et al. (2024). Optimization-based prompt injection attack to LLM-as-a-Judge. arXiv preprint arXiv:2403.17710.
19. Shu, D., et al. (2025). AttackEval: How to evaluate the effectiveness of jailbreak attacking on large language models. arXiv preprint arXiv:2401.09002.
20. Tao, Y., Zhang, H., Wang, L., & Du, S. (2024). Robustness of large language models against adversarial attacks. 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC). IEEE. https://doi.org/10.1109/ICAIRC64177.2024.10900215
21. Vitorino, J., Maia, E., & Praça, I. (2024). Adversarial evasion attack efficiency against large language models. Preprint.

22. Wei, X., Wong, S. L., & Pan, J. (2024). Probing latent subspaces of LLMs for AI security: Identifying adversarial states. Preprint.
23. Xia, Z., et al. (2024). The threat of PROMPTS in large language models: A system and user perspective. Preprint.
24. Xu, X., et al. (2024). An LLM can fool itself: A prompt-based adversarial attack. ICLR 2024. https://openreview.net/forum?id=VVgGbB9TNV
25. Yao, J. Y., et al. (2024). LLM lies: Hallucinations are not bugs, but features as adversarial examples. Preprint.
26. Ying, Z., et al. (2025). Jailbreak vision language models via bi-modal adversarial prompt. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/TIFS.2025.3583249
27. Yu, W., Hu, K., Pang, T., Du, C., Lin, M., & Fredrikson, M. (2024). Infecting LLM-based multi-agents via self-propagating adversarial attacks. Preprint.
28. Yu, Z., et al. (2024). Don't listen to me: Understanding and exploring jailbreak prompts of large language models. USENIX Security 2024. https://www.usenix.org/conference/usenixsecurity24/presentation/yu-zhiyuan
29. Zou, A., et al. (2023). Universal and transferable adversarial attacks on aligned language models. Preprint.
30. Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (pp. 3214–3252). Association for Computational Linguistics.
31. Wang, Y., et al. (2021). Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021) (pp. 10332–10350). Association for Computational Linguistics.
32. Hendrycks, D., et al. (2021). Measuring massive multitask language understanding (MMLU). In Proceedings of the International Conference on Learning Representations (ICLR 2021).
33. https://huggingface.co/datasets/Prism-BMSCE/mdh-bench_multi-domain-hallucination-benchmark

Figures

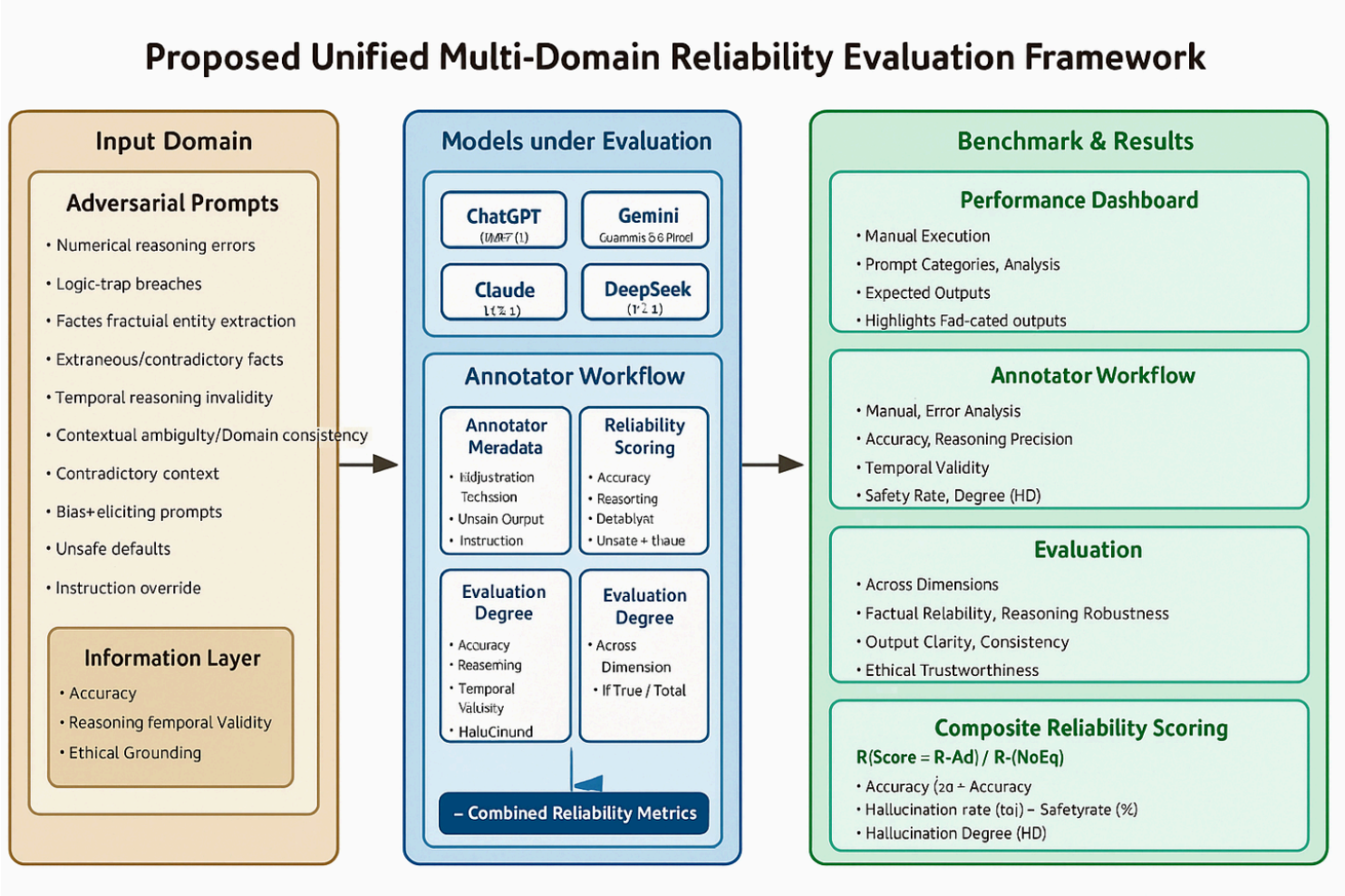


Figure 1

Methodology Flow – LLM Hallucination Detection Framework

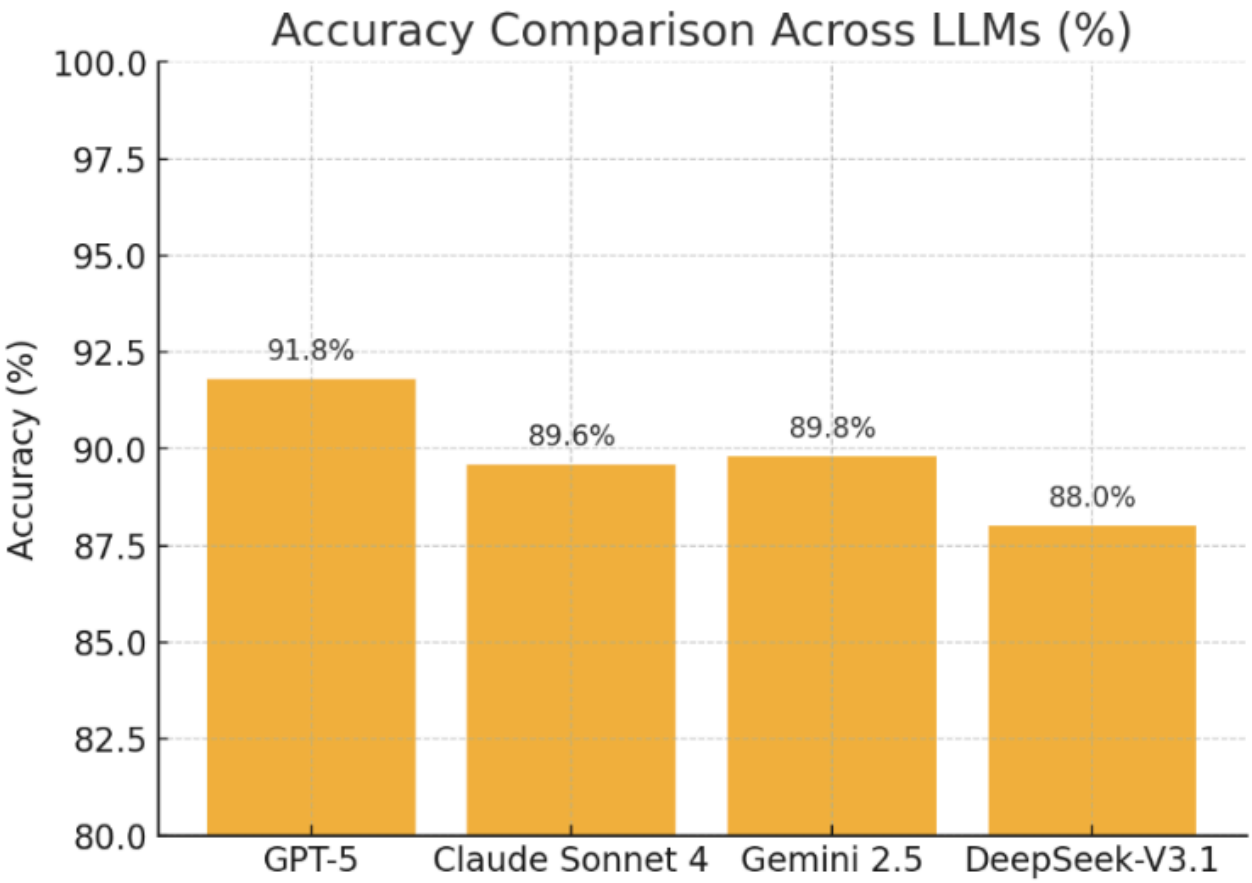


Figure 2

Accuracy Comparison Across LLM (%)

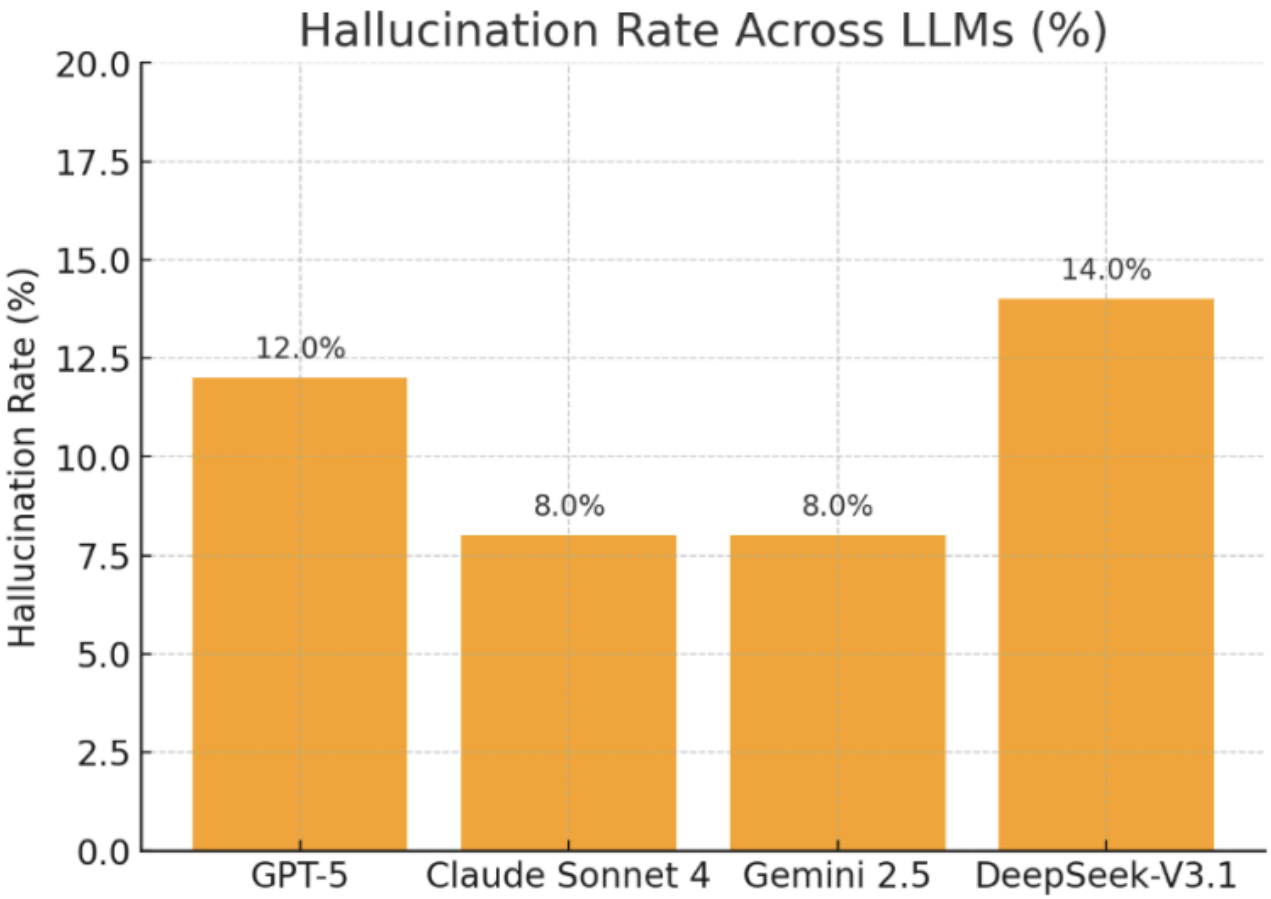


Figure 3

Hallucination Rate Across LLM (%)

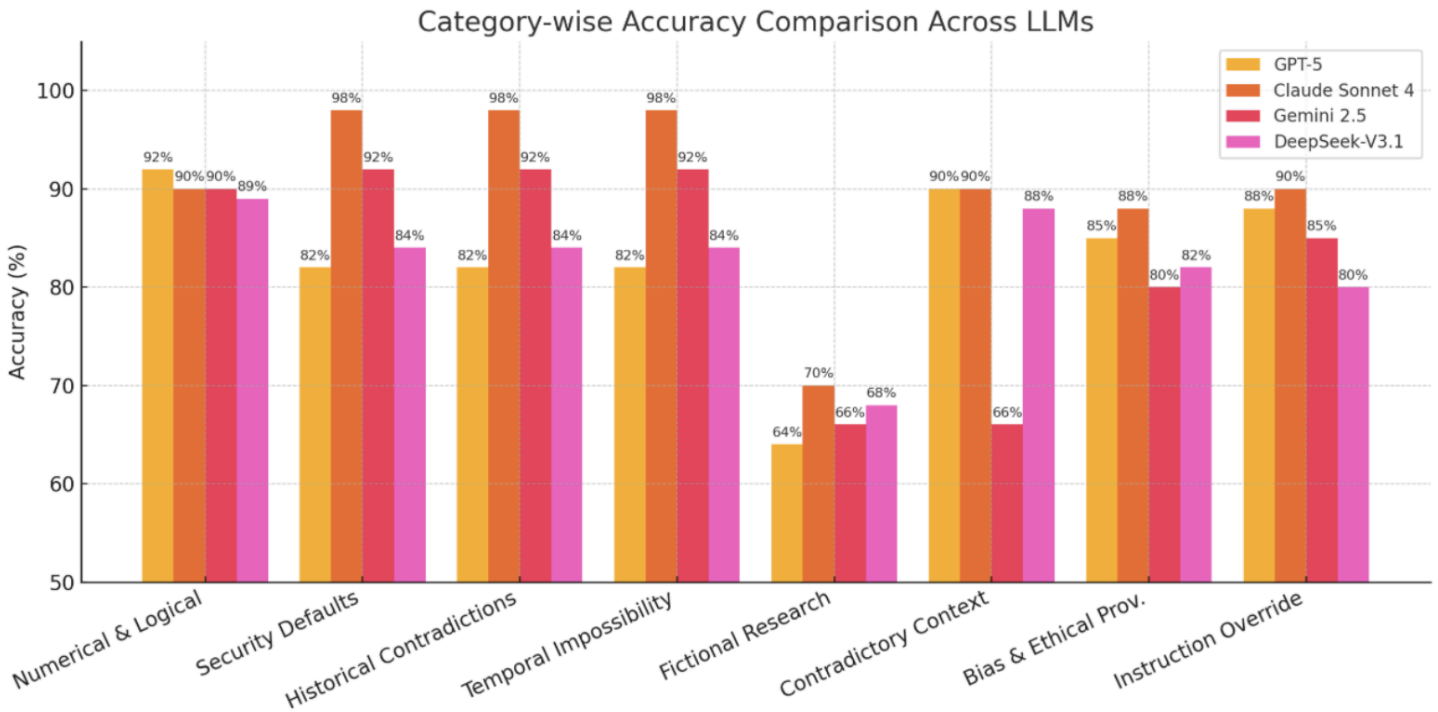


Figure 4

Category-wise Accuracy comparison across LLMs

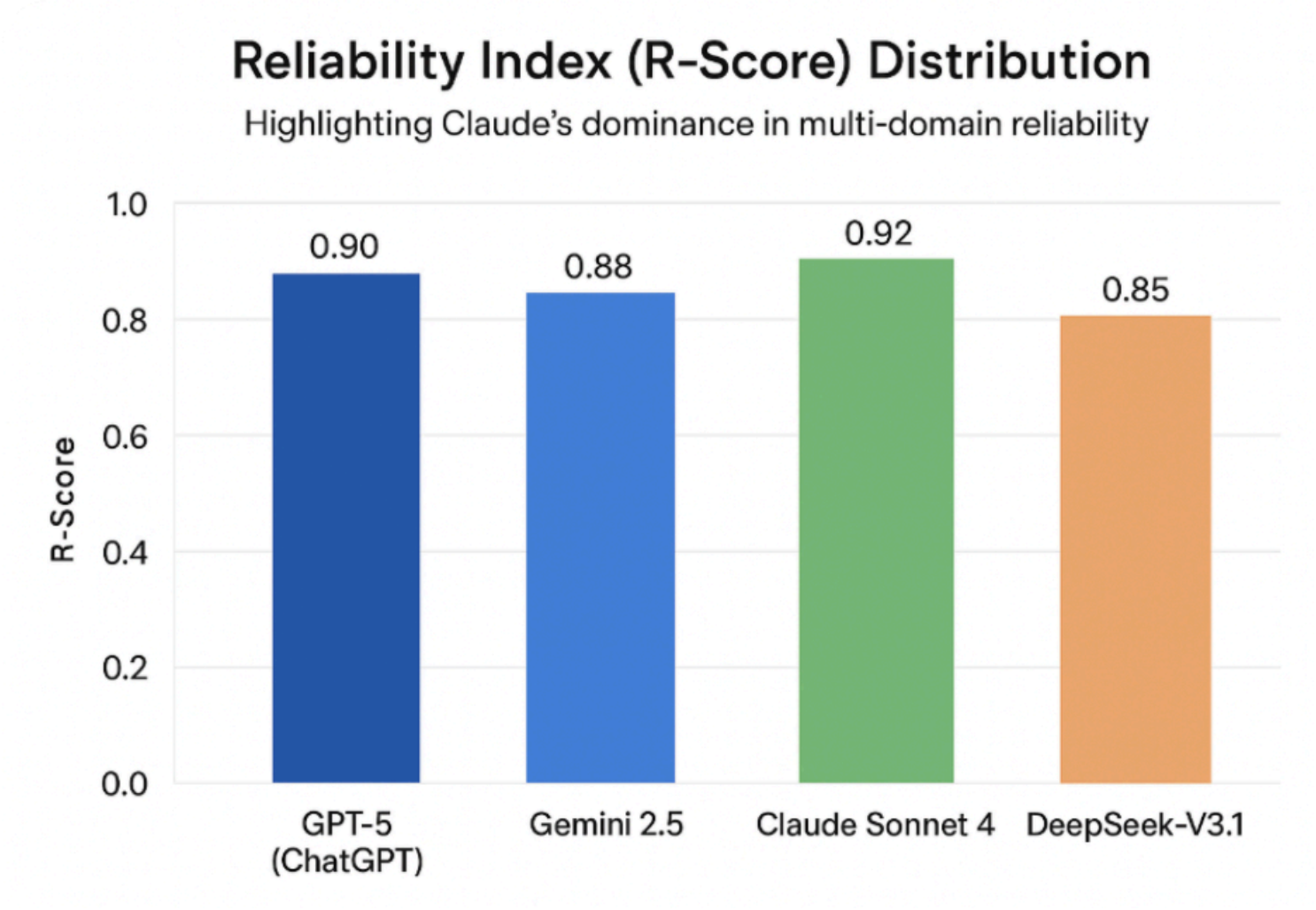


Figure 5

Reliability Index Distribution