Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081260

ARTICLE

HalluBench: A Multi-LLM Benchmark for Hallucination Evaluation and Reliability Analysis

Betül Şenyayla[1] **and Aytuğ Onan**[2,*]

[1]Department of Software Engineering, Faculty of Engineering, Sivas Cumhuriyet University, Sivas, Türkiye
[2]Department of Computer Engineering, Faculty of Engineering, İzmir Institute of Technology, İzmir, Türkiye

*Corresponding Author: Aytuğ Onan. Email: aytugonan@iyte.edu.tr

Received: 26 February 2026; Accepted: 12 May 2026; Published: 23 July 2026

ABSTRACT: Large Language Models (LLMs) have become a cornerstone of modern natural language processing, achieving strong performance across diverse tasks. Despite these advances, their tendency to generate hallucinated or factually unsupported content remains a critical challenge for reliable deployment. Existing evaluation approaches predominantly rely on single-task settings and aggregate performance metrics, implicitly assuming that hallucination behavior is uniform across tasks. However, this assumption is fundamentally flawed, as hallucination characteristics vary significantly depending on task formulation, linguistic context, and evaluation criteria. To address these limitations, this paper proposes HalluBench, a task-aware multi-LLM benchmarking framework designed for systematic hallucination analysis and metric-task alignment. The framework evaluates ten language models across four representative task formulations—open-domain question answering, cross-lingual question answering, scientific claim verification, and LLM-as-a-judge assessment—using four benchmark datasets (five evaluation splits) and nine complementary evaluation metrics. Unlike conventional approaches, HalluBench introduces a metric–task alignment strategy that selects evaluation metrics based on their suitability for each task. Experimental results reveal that hallucination behavior is strongly task-dependent, with substantial variations observed across models and evaluation settings. Specifically, the proposed framework demonstrates that model reliability is highly sensitive to task formulation; for instance, in adversarial open-domain settings, performance differences of up to 15% in Exact Match (EM) and 20% in F1 scores are observed between top-tier and compact (~1B parameter) models. By integrating lexical, semantic, and reference-based metrics within a pipeline, HalluBench provides a more robust and diagnostically informative evaluation framework compared to traditional single-task and single-metric benchmarks.

KEYWORDS: Hallucination evaluation; large language models; benchmarking framework; factual consistency; LLM reliability

1 Introduction

Large language models (LLMs) such as GPT-4, Claude, and Gemini have demonstrated remarkable performance across a wide range of natural language processing (NLP) tasks, including question answering, text summarization, reasoning, and dialogue generation [1,2]. Trained on massive and diverse corpora, these models are capable of producing fluent, coherent, and human-like responses. However, despite these advances, a fundamental limitation persists: hallucination, where models generate content that is factually incorrect, logically inconsistent, or plausibly articulated but unsupported by evidence [3]. Such hallucinated outputs undermine user trust and pose serious risks in high-stakes domains such as healthcare, education, law, and scientific research [4].

Prior work has proposed various strategies to mitigate hallucinations, including fine-tuning on curated factual datasets, reinforcement learning from human feedback, and post-hoc factual verification or correction mechanisms [5,6]. Nevertheless, evaluating hallucination behavior remains a fundamentally challenging and fragmented problem. Existing benchmarks and evaluation protocols often focus on a single task, dataset, or model family, implicitly assuming that hallucination behavior is consistent across tasks. This assumption is limiting, as hallucination characteristics vary significantly depending on task formulation, linguistic context, and evaluation methodology [7,8]. Consequently, current approaches fail to provide a unified and systematic understanding of how hallucinations vary across tasks, languages, and model architectures, as well as how such variations interact with evaluation metrics and experimental design choices [9].

While existing large-scale evaluation frameworks such as HELM [1] and BIG-Bench [2] provide broad capability coverage across diverse tasks, they primarily report aggregate performance using heterogeneous metrics without systematically analyzing how metric sensitivity and reliability vary by task formulation. As a result, hallucination robustness is often interpreted as a uniform model capability rather than a task-dependent phenomenon shaped by evaluation design.

To address these limitations, we introduce *HalluBench*, a task-aware multi-LLM benchmarking framework designed for systematic hallucination analysis and model reliability assessment. HalluBench analyzes hallucination tendencies across ten language models and four benchmark datasets spanning open-domain question answering, multilingual settings, scientific claim verification, and LLM-as-a-judge evaluation. Unlike conventional approaches, the proposed framework adopts a task-aware evaluation perspective and introduces a metric–task alignment strategy that explicitly selects evaluation metrics based on their suitability for each task. This formulation treats hallucination evaluation as a structured diagnostic problem rather than a single-score benchmarking task. Our empirical analysis reveals that model reliability is highly sensitive to task formulation; for instance, in adversarial settings, substantial performance differences are observed between large and compact models, particularly in Exact Match (EM) and F1 scores.

HalluBench further enables systematic cross-model and cross-task analyses under a unified and reproducible evaluation pipeline. Experimental findings demonstrate that hallucination behavior is inherently task-dependent and that model reliability assessments vary substantially across evaluation settings.

The main objective of this work is to establish a task-aware evaluation paradigm for hallucination analysis that improves the reliability, interpretability, and methodological consistency of LLM benchmarking. In this context, the key contributions of this paper are summarized as follows:

- We introduce a task-aware hallucination evaluation principle that demonstrates how metric–task alignment fundamentally influences reliability assessment, model ranking, and interpretability in LLM evaluation.
- We propose HalluBench, a structured multi-LLM benchmarking framework that enables systematic cross-model and cross-task analysis under controlled experimental conditions.
- We conduct a comprehensive empirical study across ten large language models and four representative task formulations using four benchmark datasets (five evaluation splits) and nine complementary evaluation metrics, revealing task-dependent hallucination dynamics.
- We provide a unified and reproducible evaluation pipeline that integrates quantitative multi-metric benchmarking with qualitative failure-mode analysis, supporting diagnostic rather than leaderboard-style evaluation.

The remainder of this paper is organized as follows. Section 2 presents the research objectives of this study. Section 3 reviews related work in hallucination evaluation. Section 4 introduces the materials used in this study, including the benchmark datasets, evaluated LLM models, and data processing and evaluation

setup. Section 5 describes the methodological framework, covering the inference strategy, task-aware metric alignment, and evaluation pipeline. Section 6 presents the proposed HalluBench framework, including the system model, architecture, and algorithmic workflow. Section 7 reports the experimentation, results, and analysis, including the experimental setup and task-specific performance evaluations. Finally, Section 8 concludes the paper by discussing the limitations and future research directions.

2 Research Objectives

The primary goal of this study is to systematically characterize hallucination behavior in large language models through a task-aware, multi-LLM evaluation framework. Rather than evaluating models in isolation, this work investigates how hallucination manifests across diverse task formulations, datasets, and evaluation metrics.

The specific research objectives are as follows:

- To analyze how hallucination behavior varies across four representative LLM task formulations: adversarial open-domain question answering, cross-lingual question answering, scientific claim verification, and LLM-as-a-judge evaluation.
- To conduct a comparative analysis of ten large language models to identify task-dependent reliability patterns and cross-model performance differences.
- To examine the effectiveness of nine complementary evaluation metrics and determine how metric–task alignment influences hallucination detection.
- To investigate recurring hallucination failure modes through both quantitative benchmarking and qualitative analysis.
- To establish a reproducible and task-aware benchmarking protocol for systematic evaluation across heterogeneous LLM settings.

This formulation positions HalluBench as both a diagnostic evaluation framework and a structured benchmark for analyzing hallucination behavior in contemporary large language models.

3 Related Works

Hallucination has emerged as a central research challenge in the development of reliable large language models (LLMs). Early investigations focused on identifying factual inconsistencies in abstractive summarization and question answering systems. For example, Kryściński et al. [10] analyzed factual consistency failures in neural summarization, while Maynez et al. [5] and Bommasani et al. [11] distinguished between intrinsic hallucinations (content contradicting the source text) and extrinsic hallucinations (content unrelated to any evidence). These insights motivated subsequent work exploring hallucination across broader task settings, including dialogue, scientific reasoning, and open-domain generation.

A growing body of literature has examined hallucination characteristics systematically. Ji et al. [3] provided a comprehensive survey categorizing hallucination sources, linguistic triggers, and evaluation methodologies. Bang et al. [9] conducted a large-scale multitask evaluation of ChatGPT across reasoning, hallucination, and multilingual scenarios, highlighting persistent failure modes even in advanced models. More recently, Bang et al. [12] introduced HalluLens, a benchmark for hallucination detection and categorization, offering taxonomy-driven evaluations and qualitative error analysis for multiple instruction-tuned models.

To quantify hallucination behavior, several metrics and evaluation frameworks have been proposed. Laban et al. [8] introduced faithfulness scoring for summarization, while Min et al. [6] developed FActScore

to assess factual consistency using entity-level grounding. Li et al. [7] proposed HaluEval, a large-scale benchmark using reference-based scoring for multiple generative tasks. While these metric-driven approaches contribute to measuring hallucination severity, they are often applied uniformly across tasks, despite growing evidence that evaluation reliability is strongly task-dependent. Recent studies demonstrate that no single metric or evaluator generalizes reliably across different task formulations, output structures, and reference settings [13,14]. In addition, benchmark design efforts emphasize that misalignment between task definitions and evaluation metrics can lead to unstable or misleading performance conclusions [15].

Benchmarking efforts have expanded to evaluate generative reliability in broader contexts. Datasets such as TruthfulQA [4], FEVER [16], and XSumFaith [17] provide factuality assessments for question answering and summarization tasks. HELM [1] and BIG-Bench [2] represent comprehensive large-scale evaluations that incorporate multiple models, datasets, and prompt types, offering insights into model behavior beyond single-task performance. Liu et al. [18] further explored structured output reliability, demonstrating that hallucination and factual correctness issues extend to constrained text generation tasks.

In parallel, ensemble-based and voting-based approaches have been widely explored to improve aggregate model performance by combining outputs from multiple models or agents. For instance, Zhu et al. [19] propose voting-based aggregation strategies in the context of differentially private federated learning, showing that collective decision mechanisms can enhance robustness and stability. However, such approaches aim to optimize final predictions through aggregation and do not explicitly analyze hallucination behavior across independently generated model outputs. A comparative multi-LLM evaluation perspective, in contrast, enables direct analysis of consistency, robustness, and failure modes across individual models rather than improving performance via ensemble mechanisms.

Recent work has further emphasized the importance of qualitative failure analysis for understanding the limitations of LLMs. Stewart et al. [20] investigate failure mode classification in LLM outputs, demonstrating that errors often manifest as distinct qualitative patterns such as under-informative responses, unsupported claims, or misaligned outputs. Complementing this view, Vinay [21] proposes a system-level taxonomy of failure modes in LLM-based applications, highlighting that reliability issues extend beyond isolated factual errors to broader behavioral and structural limitations. Despite these advances, most existing benchmarks continue to prioritize aggregate quantitative scores, offering limited support for fine-grained qualitative analysis of hallucination behaviors across models and tasks.

Despite the substantial progress in hallucination analysis and evaluation, several critical technical gaps remain insufficiently addressed in the literature. First, most current studies focus on either single-model evaluation or task-specific benchmarks, which fails to capture cross-model consistency and divergent failure modes across heterogeneous task formulations [3,22]. Second, a major technical limitation is the "one-size-fits-all" use of evaluation metrics; existing frameworks predominantly apply aggregate scoring (e.g., Accuracy, F1) without considering how metric sensitivity and reliability vary depending on task structure, leading to potentially misleading model rankings [15].

Furthermore, there is a lack of principled mechanisms to align evaluation criteria with linguistic and structural requirements across open-domain, cross-lingual, and scientific settings, leaving a gap in understanding how hallucination behaviors vary by context. The interaction between evaluation metrics and task characteristics remains underexplored, despite growing evidence that metric sensitivity depends strongly on task formulation and output structure. In addition, cross-lingual and multilingual evaluation scenarios are often treated as extensions of monolingual tasks, without adequately addressing linguistic variability and its impact on hallucination assessment.

Another limitation lies in the lack of unified and reproducible evaluation pipelines that integrate multiple datasets, models, and metrics under consistent experimental conditions. Existing frameworks often differ in preprocessing steps, prompt design, and evaluation configurations, making systematic comparison across studies challenging. Moreover, qualitative analysis of hallucination patterns is typically conducted in an ad hoc manner, rather than being systematically integrated into benchmarking workflows.

In addition to these approaches, several recent studies have explored hallucination evaluation through multi-model and meta-evaluation perspectives. For example, JudgeBench [23] introduces an LLM-as-a-judge framework to assess consistency and reliability across model-generated outputs, highlighting the variability of evaluation outcomes depending on the chosen evaluator. Similarly, recent benchmarking efforts emphasize the importance of combining multiple evaluation signals and moving beyond single-metric assessment [6–8]. However, these approaches often focus on specific task settings or evaluation paradigms, without explicitly modeling the relationship between task formulation and metric sensitivity. In contrast, the proposed HalluBench framework extends this line of work by introducing a task-aware metric alignment strategy within a unified multi-LLM evaluation pipeline.

To address these limitations, the present work introduces HalluBench, a task-aware multi-LLM hallucination benchmark designed to enable consistent evaluation across ten language models and four diverse datasets. By introducing a metric–task alignment strategy within a unified and reproducible pipeline, HalluBench provides a structured framework for analyzing hallucination behavior across tasks, models, and evaluation settings.

4 Materials and Methods

This section presents the materials and methodological framework adopted in this study for task-aware hallucination evaluation. Section 4.1 describes the benchmark datasets, evaluated language models, data preprocessing procedures, and feature extraction mechanisms used to ensure consistent analysis.

Section 4.2 details the underlying evaluation protocol, including inference strategy and task-aware metric alignment. Together, these components define the HalluBench framework as a structured and reproducible system for systematically analyzing hallucination behavior across heterogeneous tasks, models, and evaluation settings.

4.1 Materials

This subsection describes the data sources, evaluated models, and data handling procedures used in this study. These components form the foundation of the HalluBench framework and ensure consistent and reproducible evaluation across heterogeneous LLM settings.

4.1.1 Datasets

In this study, we employ publicly available hallucination evaluation benchmarks to establish a comprehensive framework for analyzing factual consistency in large language models (LLMs). To ensure transparency, reproducibility, and independent verification of experimental findings, the evaluation relies exclusively on well-documented and openly accessible datasets obtained from established NLP repositories. Each selected dataset captures a distinct dimension of hallucination, including open-domain factuality, cross-lingual robustness, scientific claim verification, and LLM-as-a-judge consistency, as detailed below.

These datasets were selected to represent complementary evaluation scenarios, including open-domain question answering, multilingual reasoning, evidence-based fact verification, and judgment stability. This

design enables systematic comparison of hallucination behavior across diverse task formulations while maintaining reproducible and transparent experimental conditions.

All datasets used in this study are publicly available via HuggingFace repositories.

- **TruthfulQA (Lin et al., 2022):** TruthfulQA consists of 817 manually curated questions designed to assess whether models generate truthful answers to commonly misunderstood or misconception-based queries. Each item includes a truthful reference answer along with multiple adversarial false answers, enabling systematic evaluation of factual accuracy and hallucination behavior in open-domain settings [4].
- **TruthfulQA-TR (Mukayese AI Team, 2023):** TruthfulQA-TR is the Turkish adaptation of the original TruthfulQA benchmark, translated and culturally localized to support cross-linguistic hallucination evaluation in Turkish. The dataset consists of 817 questions, mirroring the structure of the original benchmark, and reflects semantic drift and morphological challenges characteristic of low-resource language settings. It enables reproducible, license-free benchmarking of cross-lingual factual consistency [24].
- **SciFact (Wadden et al., 2020):** SciFact is a scientific claim verification dataset designed to evaluate factual consistency in evidence-based reasoning tasks. In this study, we utilize the HuggingFace *mteb/scifact* variant, which contains 7550 query–document pairs derived from the original SciFact benchmark. This formulation enables evaluation of semantic retrieval and evidence alignment, supporting analysis of hallucination behavior in scientific reasoning settings [25,26].
- **JudgeBench (Huang et al., 2024):** JudgeBench is a benchmark designed to evaluate factual reliability and reasoning consistency in multi-LLM settings using an LLM-as-a-judge paradigm. In this study, we utilize two evaluation splits, consisting of approximately 270 examples (Claude split) and 350 examples (GPT split). The benchmark comprises diverse question–answer pairs evaluated by large language models, assessing factual correctness, reasoning coherence, and response alignment. JudgeBench supports scalable meta-evaluation of hallucination by enabling analysis of cross-judge agreement and evaluation stability across generative language models [23].

Overall, the datasets employed in this study are selected to capture complementary dimensions of hallucination across diverse evaluation tasks in large language models. TruthfulQA and TruthfulQA-TR support truthfulness and factuality assessment in open-domain question answering, enabling analysis of hallucinations in both English and Turkish settings. SciFact facilitates evaluation of evidence-based reasoning through scientific claim verification, while JudgeBench enables meta-evaluation under the LLM-as-a-judge paradigm, capturing consistency and agreement across model-generated judgments.

Taken together, this dataset configuration enables a comprehensive analysis of hallucination behavior not only in terms of factual correctness, but also with respect to reasoning consistency, linguistic generalization, and evaluation stability across models. To further illustrate the diversity of task structures and hallucination risk profiles across the selected benchmarks, Table 1 presents representative examples from each dataset, while Table 2 summarizes the number of evaluation samples used in each benchmark.

4.1.2 LLM Models

To comprehensively evaluate hallucination behaviors across diverse model architectures and training paradigms, we employed several state-of-the-art LLMs developed by both leading research organizations and the open-source community. The selection criteria focused on model diversity, accessibility, and reported capabilities on reasoning- and factuality-oriented tasks.

Table 1: Representative examples illustrating different evaluation tasks across benchmark datasets (five evaluation splits from four datasets).

Dataset	Task Type	Example Input/Query	Expected Reference	Domain
TruthfulQA	Open-domain Question Answering	What is the smallest country in the world that is a member of the United Nations?	"Nauru is the smallest by land area."	General Knowledge
TruthfulQA-TR	Cross-lingual Question Answering	Dünyada yüzölçümü en az olan Birleşmiş Milletler üyesi ülke nedir?	"Nauru dünyanın en küçük yüzölçümlü ülkesidir."	General Knowledge
SciFact	Scientific Claim Verification	Bayesian measures of model uncertainty are widely applicable across biomedical image segmentation tasks.	Supported/Refuted (based on cited evidence)	Biomedical
JudgeBench (Claude split)	LLM-as-a-Judge (Pairwise Preference)	In the brain stem, which structure controls involuntary breathing?	A > B (gold: A preferred)	Psychology
JudgeBench (GPT split)	LLM-as-a-Judge (Pairwise Preference)	Which ethical principle best applies to a ceramics studio environmental policy?	B > A (gold: B preferred)	Law

Note: A > B indicates the model output A is preferred over B according to the gold label.

Table 2: Number of evaluation samples used per dataset.

Dataset	Number of Samples
TruthfulQA	817 questions
TruthfulQA-TR	817 questions
SciFact (MTEB)	7550 samples
JudgeBench (Claude split)	270 examples
JudgeBench (GPT split)	350 examples

Note: All available samples from each dataset were used without sub-sampling. Each model was evaluated on the same set of inputs to ensure fair and consistent comparison.

The evaluated systems span multiple categories, including commercial proprietary models, instruction-tuned open-weight models, compact reasoning models, and multimodal variants, enabling a balanced and task-aware analysis of how different design choices influence hallucination tendencies. Model sizes in our evaluation range from compact 1B-scale architectures to commercial systems exceeding tens of billions of parameters, supporting systematic comparison across model scales.

Proprietary models were accessed via official API endpoints, whereas open-weight systems were executed locally in a controlled environment. For readability, Table 3 presents an overview of key model characteristics, with detailed descriptions provided afterward. In total, 10 LLMs were evaluated.

Table 3: Overview of evaluated large language models used in this study.

Model	Type	Params	Access	Key Characteristics
GPT-4o-2024-05-13	Proprietary	Proprietary (est. >50B)	API	Multimodal reasoning, high factual reliability
Claude-3.5	Proprietary	Proprietary (est. >50B)	API	Strong coherence, robust long-form consistency
Phi-4-Mini-Reasoning	Open-weight	<4B	Local	Instruction-tuned compact model optimized for reasoning
Llama-3.2-1B	Open-weight	1B	Local	Lightweight model for low-resource hallucination tests
Mistral-7B-Instruct-v0.3	Open-weight	7B	Local	Instruction-tuned chat model with stable outputs
Falcon-RW-1B	Open-weight	1B	Local	Efficient decoder-only architecture
Gemma-7B	Open-weight	7B	Local	Google-released multilingual LLM
DeepSeek-R1-Distill-Qwen-1.5B	Open-weight	1.5B	Local	Distilled reasoning model based on Qwen architecture
OpenChat-3.5-1210	Open-weight	7B	Local	Chat-optimized model with strong short-form responses
C4AI-Command-R7B	Open-weight	7B	Local	RAG-friendly model with reliable factual grounding

Note: Proprietary models were accessed via official APIs, whereas open-weight models were executed locally in a controlled environment to support reproducibility.

- **GPT-4o (OpenAI, 2024):** GPT-4o is a proprietary multimodal large language model designed for real-time interaction and low-latency generation. The model is accessible exclusively via an API, as its architecture and model weights are not publicly disclosed. GPT-4o supports instruction-following and long-form text generation, making it suitable for analyzing hallucination behavior in interactive and information-seeking scenarios [27].
- **Claude 3.5 Sonnet (Anthropic, 2024):** Claude 3.5 Sonnet is a proprietary large language model developed by Anthropic as part of the Claude 3 family. It supports instruction-following and long-form text generation and is accessible exclusively via API, as its model weights and architectural details are

not publicly disclosed. Claude 3.5 Sonnet serves as a commercial baseline for evaluating hallucination tendencies in dialogue and reasoning-oriented tasks [28].

- **Command R7B (CohereLabs, 2024):** Command R7B is a 7B-parameter, instruction-tuned open-weight language model developed by CohereLabs. It supports reproducible local deployment, making it a practical open-source baseline for evaluating hallucination behavior in dialogue-oriented and controlled generation settings [29].
- **Llama-3.2–1B (Meta, 2024):** LLaMA-3.2–1B is a lightweight, open-weight language model with approximately 1B parameters, designed for efficient inference in low-resource settings. Its compact architecture enables fast local inference and serves as a reproducible baseline for examining the relationship between parameter scale, factual consistency, and hallucination tendencies [30].
- **Falcon-RW-1B (TII, 2023):** Falcon-RW-1B is a 1B-parameter, open-weight autoregressive transformer language model released by the Technology Innovation Institute (TII). Its efficient inference profile makes it suitable as a reproducible small-scale baseline for examining hallucination tendencies in resource-limited settings [31].
- **Gemma-7B (Google, 2024):** Gemma-7B is an open-weight language model released by Google as part of the Gemma family, with approximately 7B parameters. It supports general-purpose text generation and reproducible experimentation, serving as a mid-scale open-weight baseline for evaluating hallucination behavior [32].
- **DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek, 2024):** DeepSeek-R1-Distill-Qwen-1.5B is a compact, open-weight language model derived from the Qwen architecture with approximately 1.5B parameters. As a distilled model, it serves as a lightweight baseline for examining hallucination behavior and scale-related effects in resource-constrained settings [33].
- **OpenChat-3.5-1210 (OpenChat, 2023):** OpenChat-3.5 is an open-weight, instruction-tuned conversational language model designed for multi-turn dialogue. It serves as a publicly accessible baseline for analyzing hallucination behavior in dialogue-oriented generation and comparison with proprietary models [34].
- **Phi-4-Mini-Reasoning (Microsoft, 2024):** Phi-4 Mini Reasoning is a compact, open-weight language model with approximately 3.8B parameters. Its lightweight architecture makes it suitable as a baseline for examining hallucination behavior and parameter-efficiency relationships in smaller-scale reasoning models [35].
- **Mistral-7B-Instruct-v0.3 (Mistral AI, 2024):** Mistral-7B-Instruct-v0.3 is an open-weight, instruction-tuned language model with approximately 7B parameters. It serves as a widely adopted baseline for hallucination evaluation and reproducible comparison with larger proprietary language models [36].

Together, these models form a diverse and systematic evaluation suite for hallucination benchmarking. This configuration enables comparative analysis of the relationship between parameter scale, training strategy, alignment objectives, and factual consistency and reasoning stability in large language models.

4.1.3 Data Processing and Evaluation Setup

In this study, the evaluation process is guided by a task-aware metric selection strategy, where primary metrics are selected according to the functional requirements of each task. Specifically, EM and F1 are prioritized for open-domain and cross-lingual question answering, whereas BERTScore and BLEURT are emphasized for scientific claim verification and LLM-as-a-judge evaluation. Remaining metrics are reported as secondary indicators to provide complementary diagnostic signals and to support a more comprehensive interpretation of model behavior.

Primary metrics are selected based on their direct alignment with task objectives, such as strict factual correctness in question answering or semantic consistency in reasoning tasks. Secondary metrics are included to provide complementary signals, capturing additional aspects such as lexical variation, structural similarity, and partial semantic alignment, which may not be fully reflected by primary metrics alone.

For each dataset, all entries were processed without sub-sampling, preserving the original distribution of topics and difficulty levels. Each sample was paired with model-generated outputs and analyzed using predefined hallucination categories, including (i) factual error, (ii) unsupported claim, (iii) fabricated detail, (iv) reasoning inconsistency, and (v) irrelevant response. These categories were used for qualitative analysis and error characterization rather than direct metric-based scoring.

Quantitative evaluation was conducted using a diverse collection of automatic metrics, including:

- lexical overlap metrics (Exact Match (EM), F1, BLEU, ROUGE-1/2/L),
- semantic similarity metrics (METEOR, BERTScore),
- reference-based contextual fidelity metrics (BLEURT).

These metrics collectively capture lexical accuracy, semantic alignment, and reference-based consistency, enabling fine-grained analysis of hallucination behavior across tasks. Representative task–dataset mappings and example inputs are summarized in Table 1.

Prior to metric computation, model outputs and reference answers were normalized using lowercasing, punctuation removal, and whitespace normalization. Token-level metrics such as Exact Match and F1 were computed using whitespace-based tokenization without applying language-specific stemming or morphological normalization. This preprocessing reduces formatting-related variance while preserving surface-form differences, enabling consistent comparison across English and Turkish evaluation settings.

Turkish characters were intentionally preserved during normalization as part of a language-aware evaluation design, preventing artificial lexical convergence caused by character folding [37]. Given the agglutinative morphology of Turkish, avoiding additional stemming or morphological normalization helps maintain faithful surface-form comparison and prevents inflated lexical overlap scores.

Each evaluation metric was computed using its recommended preprocessing configuration to preserve methodological validity and ensure consistent and reliable evaluation across heterogeneous tasks and datasets.

4.1.4 Augmentation

No explicit data augmentation techniques were applied in this study. The evaluation relies on original benchmark datasets in order to preserve the natural distribution of hallucination patterns and avoid introducing artificial biases into the evaluation process.

4.2 Methods

This study is grounded in a task-aware evaluation paradigm that treats hallucination not as a uniform model property, but as a context-dependent phenomenon influenced by task formulation, linguistic structure, and evaluation criteria. The proposed methodology builds upon established evaluation practices in natural language generation, including lexical overlap metrics, embedding-based similarity measures, and reference-based evaluation techniques, which have been widely used for assessing factual consistency and hallucination in prior work.

The framework extends these approaches by introducing a task-aware metric alignment strategy and a structured multi-LLM evaluation design. It adopts a multi-metric and diagnostic evaluation strategy,

where different metrics capture complementary aspects of hallucination behavior. This design is further supported by a comparative multi-LLM analysis framework, enabling systematic investigation of consistency, robustness, and failure modes across models. Accordingly, this subsection describes the methodological design of the proposed HalluBench framework, focusing on the evaluation pipeline, inference strategy, and task-aware metric alignment.

4.2.1 Inference Strategy

To ensure comparability, all models were evaluated using a unified prompt template and deterministic decoding settings. Open-weight models were executed with temperature = 0.0 and max_tokens = 512. Proprietary API-based systems (GPT-4o and Claude-3.5) were run using equivalent zero-temperature or greedy decoding modes when available. For open-weight models, all random processes were controlled using a fixed seed (42) to promote deterministic execution and reproducible results.

These parameter settings are selected to minimize stochastic variation in model outputs, thereby enabling fair, controlled, and reproducible comparison across models. In particular, the use of zero temperature enforces deterministic generation, eliminating randomness in token selection and ensuring consistency across repeated runs. The fixed random seed further stabilizes the generation process for open-weight models.

The chosen maximum token limit (512) provides sufficient capacity for complete and contextually adequate responses across all tasks, while preventing excessively long outputs that may introduce noise or bias into evaluation metrics. This balance ensures that model outputs remain comparable in length and structure, supporting consistent and reliable metric-based evaluation.

4.2.2 Task-Aware Metric Alignment

The proposed framework adopts a task-aware metric alignment strategy, where evaluation metrics are selected based on the functional requirements of each task rather than being applied uniformly. For question answering tasks, Exact Match (EM) and F1 are treated as primary indicators of factual correctness. In contrast, for tasks requiring semantic reasoning, such as scientific claim verification and LLM-as-a-judge evaluation, embedding-based metrics such as BERTScore and BLEURT are prioritized.

This design enables more reliable and context-sensitive evaluation by aligning metric behavior with task characteristics.

4.2.3 Evaluation Pipeline

The proposed evaluation pipeline follows a standardized and reproducible process that integrates dataset processing, prompt construction, controlled model inference, output normalization, and multi-metric evaluation within a unified framework.

Each input instance is processed using a unified prompt template to ensure consistency across models. Model outputs are generated under deterministic decoding configurations, eliminating stochastic variation and ensuring reproducibility. The generated outputs are subsequently collected and normalized using standard preprocessing steps, including lowercasing, punctuation removal, and whitespace normalization.

Following normalization, outputs are evaluated using a diverse set of complementary metrics, including lexical overlap, semantic similarity, and reference-based evaluation measures. Finally, the resulting scores are aggregated at the dataset level to enable systematic comparison of hallucination behavior across different models, datasets, and task formulations.

This structured pipeline ensures that all models are evaluated under identical experimental conditions, supporting fair, consistent, and reproducible analysis across heterogeneous LLM settings.

5 Proposed Framework: HalluBench

To address this challenge, we propose *HalluBench*, a task-aware multi-LLM benchmarking framework designed to evaluate hallucination behavior across heterogeneous evaluation settings. The framework integrates multiple datasets, models, and evaluation metrics into a unified pipeline, enabling systematic analysis of factual reliability.

The core novelty of HalluBench lies in its task-aware metric alignment mechanism, which explicitly models the interaction between task structure and evaluation metrics, enabling more reliable and interpretable hallucination analysis compared to traditional aggregate evaluation approaches.

5.1 System Model

We formally define the HalluBench framework as a structured evaluation system consisting of datasets, models, and evaluation metrics.

Let:

- $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ denote a dataset of input–reference pairs,
- $\mathcal{M} = \{M_1, M_2, ..., M_K\}$ denote a set of evaluated language models,
- $\hat{y}_i^{(k)} = M_k(x_i)$ denote the output generated by model M_k for input x_i,
- $\mathcal{F} = \{f_1, f_2, ..., f_L\}$ denote a set of evaluation metrics.

For each model M_k and dataset $\mathcal{D}$, hallucination behavior is evaluated by computing:

$$S_{k,l} = \frac{1}{N} \sum_{i=1}^{N} f_l \left(\hat{y}_i^{(k)}, y_i \right)$$

where $S_{k,l}$ represents the score of model M_k under metric f_l.

Unlike traditional evaluation approaches that aggregate all metrics into a single score, HalluBench introduces a task-aware metric alignment function:

$$\mathcal{A} : \mathcal{T} \rightarrow \mathcal{F}$$

which maps each task type $\mathcal{T}$ to a subset of relevant evaluation metrics.

Thus, the final evaluation is defined as:

$$S_k^{task} = \{S_{k,l} \mid f_l \in \mathcal{A}(task)\}$$

This formulation enables task-specific evaluation by conditioning metric selection on task characteristics, thereby avoiding misleading conclusions associated with uniform metric aggregation and improving the reliability and interpretability of hallucination assessment across heterogeneous evaluation settings.

5.2 Architecture and Working

The HalluBench framework follows a modular and task-aware pipeline architecture designed to systematically capture hallucination behavior across heterogeneous evaluation settings. This pipeline explicitly integrates task-aware metric selection within a multi-LLM evaluation setting, distinguishing HalluBench from conventional evaluation workflows that rely on uniform metric aggregation.

The overall workflow of the proposed framework is defined as follows:

1. **Input Acquisition:** Input samples are retrieved from benchmark datasets.
2. **Prompt Construction:** A unified and task-consistent prompt template is applied to ensure comparability across models.
3. **Model Inference:** Each model generates outputs under deterministic decoding settings to ensure reproducibility and eliminate stochastic variation.
4. **Output Normalization:** Generated outputs and reference answers are normalized.
5. **Metric Evaluation:** Multiple complementary evaluation metrics are computed to capture lexical, semantic, and contextual aspects of hallucination behavior.
6. **Task-aware Aggregation:** Metrics are selected based on task formulation.
7. **Result Analysis:** Quantitative and qualitative hallucination analysis is performed.

5.2.1 Algorithm

The overall evaluation workflow of the proposed HalluBench framework can be formally summarized as follows:

1. For each dataset $\mathcal{D}$
2. For each model $M_k \in \mathcal{M}$
   (a) Generate output $\hat{y}_i^{(k)} = M_k(x_i)$
   (b) Normalize outputs and references
   (c) Compute all metrics $f_l(\hat{y}_i^{(k)}, y_i)$
   (d) Select task-relevant metrics using $\mathcal{A}$
3. Aggregate, store, and analyze results

As illustrated in Fig. 1, the proposed HalluBench framework follows a structured pipeline from prompt construction to hallucination analysis, integrating task-aware metric selection within a multi-LLM evaluation setting.

6 Experimantation, Results and Analysis

This section presents the benchmarking experiments conducted using the proposed HalluBench framework to systematically evaluate hallucination behavior across multiple large language models. The experimental design follows a task-aware evaluation paradigm, enabling comparative analysis of model reliability, task-dependent performance variation, and metric sensitivity across heterogeneous evaluation settings. The section further details the experimental setup, evaluation metrics, and quantitative and qualitative results.

6.1 Experimental Setup

All experiments were conducted in a cloud-based Python environment using Google Colab Pro. Open-weight models were executed using GPU acceleration (primarily NVIDIA T4, depending on session allocation) with PyTorch and HuggingFace Transformers libraries. Proprietary models were accessed via their official API endpoints under equivalent inference configurations.

To ensure consistency and comparability across experiments, a unified prompt template was applied to all models. Consistent with the inference strategy described in Section 4.2.1, deterministic decoding settings (temperature = 0.0, max_tokens = 512) and fixed random seeds (42) were used for all applicable models to minimize stochastic variation and ensure reproducibility.

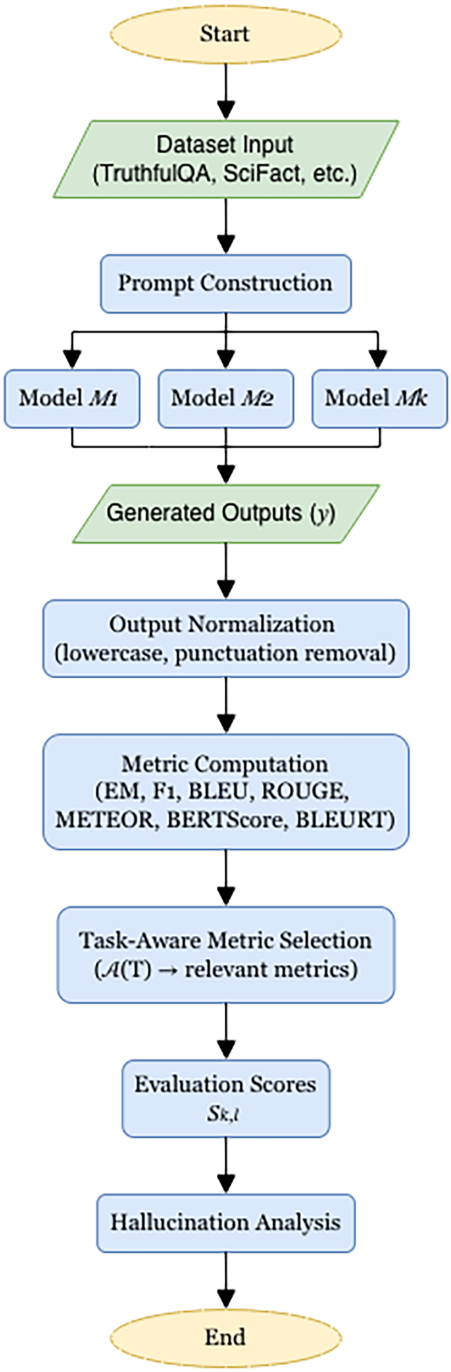

Figure 1: Overall workflow of the proposed HalluBench framework, illustrating the evaluation pipeline from prompt construction to hallucination analysis, with task-aware metric selection.

All datasets were evaluated using their full splits without sub-sampling, and each model was tested on the same set of inputs under identical experimental conditions.

6.2 Reproducibility Statement

To ensure scientific rigor and facilitate reproducibility, all experimental procedures, configuration settings, and evaluation protocols used in this study are documented. A replication package will be made publicly available upon acceptance of this paper. This package will include prompt templates, detailed model configuration settings, dataset usage specifications, and evaluation scripts to enable independent verification and reproduction of the experimental results.

All datasets employed in this study are publicly accessible via HuggingFace repositories. The provided resources are intended to enable independent verification of the reported results and to support extensibility of the proposed HalluBench framework in future research.

6.3 Evaluation Metrics

Within the proposed task-aware evaluation framework, these metrics are not applied uniformly; instead, they are selected and interpreted based on their suitability for each task formulation. Specifically, metrics are categorized into primary and secondary indicators depending on their diagnostic relevance. Primary metrics are selected based on their direct alignment with task objectives (e.g., EM and F1 for question answering tasks, BERTScore and BLEURT for evidence-based reasoning), while secondary metrics (e.g., BLEU, ROUGE, METEOR) provide complementary signals for analyzing lexical and structural similarity. The selected metrics capture both lexical overlap and semantic similarity between model-generated and reference outputs, enabling a multidimensional and comprehensive evaluation of hallucination behavior. These metrics are widely adopted in the factuality and hallucination literature and have been shown to be effective for evaluating truthfulness and consistency in generative models [3,6,10].

The Exact Match (EM) metric was used to measure the strict correctness of model predictions by verifying whether the generated response exactly matches the reference answer. EM is formally defined as:

$$EM = \frac{\text{Number of exact matches}}{\text{Total number of samples}} \tag{1}$$

Before computing EM, all predictions and reference answers were normalized through lowercasing, whitespace trimming, and punctuation removal, following standard question answering benchmark practices (e.g., SQuAD and TruthfulQA). This normalization preserves semantic content while ensuring fair comparison across models. Without normalization, superficial formatting differences such as capitalization or trailing punctuation could incorrectly penalize otherwise correct answers.

While EM provides a strict measure of factual precision, it does not assign partial credit to responses that are nearly correct. To capture partial factual alignment, the F1 score was computed by balancing precision and recall at the token level. Precision and recall are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

where TP denotes the number of overlapping tokens between the model-generated response and the reference answer, FP denotes tokens generated by the model but not present in the reference, and FN denotes reference tokens that were not captured by the model output. The F1 score complements Exact Match by

providing a balanced measure of partial factual alignment between model-generated responses and reference answers, enabling finer-grained evaluation of hallucination-related errors [38].

In addition to token-level metrics, BLEU (Bilingual Evaluation Understudy) was employed to measure n-gram overlap between model-generated and reference responses, capturing surface-level lexical similarity and phrasing consistency. BLEU is formally computed as:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5}$$

where p_n denotes the modified n-gram precision, w_n represents the weighting factors (typically uniform), and BP denotes the brevity penalty applied to penalize overly short outputs. While BLEU is effective for assessing lexical overlap, it does not directly capture semantic correctness or factual consistency, and is therefore used as a complementary metric rather than a primary indicator of hallucination behavior.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), particularly ROUGE-N, was employed to measure recall-oriented n-gram overlap between model-generated outputs and reference answers. ROUGE-N is formally defined as:

$$ROUGE\text{-}N = \frac{\sum_{\text{n-gram}\in\text{Ref}} \text{Count}_{\text{match}}(\text{n-gram})}{\sum_{\text{n-gram}\in\text{Ref}} \text{Count}(\text{n-gram})} \tag{6}$$

In our experiments, we adopt multiple ROUGE variants, including ROUGE-1 and ROUGE-2, which correspond to unigram and bigram overlap, respectively. In addition, ROUGE-L is employed to measure sequence-level similarity based on the longest common subsequence between generated and reference texts. While ROUGE metrics effectively capture recall-oriented lexical and structural overlap, they do not directly assess semantic correctness or factual grounding, and are therefore used as complementary indicators in hallucination evaluation.

This metric is especially useful for assessing longer or more descriptive outputs, where capturing all reference content is critical. BLEU emphasizes precision, whereas ROUGE focuses on recall, making them complementary for evaluating lexical overlap from different perspectives.

To further enhance evaluation sensitivity, METEOR (Metric for Evaluation of Translation with Explicit Ordering) was employed. Unlike BLEU and ROUGE, METEOR incorporates synonym matching, stemming, and word order alignment. METEOR is computed as a weighted harmonic mean of precision and recall, with an additional penalty for fragmented matches:

$$F_{\text{mean}} = \frac{10 \cdot Precision \cdot Recall}{Recall + 9 \cdot Precision} \tag{7}$$

$$METEOR = F_{\text{mean}} \cdot (1 - Penalty) \tag{8}$$

where the penalty increases when matched words are not contiguous, thereby improving sensitivity to both semantic and syntactic correctness. By incorporating synonym matching and alignment-based scoring, METEOR captures semantic fidelity beyond strict lexical overlap, making it particularly suitable for evaluating factual consistency in free-form generative outputs.

Taken together, BLEU [10], ROUGE [8], and METEOR [5] serve as lexical overlap metrics that capture surface-level similarity between generated text and reference answers. These metrics quantify how closely model outputs reproduce reference phrasing, providing interpretable indicators of hallucination-related errors in long-form or descriptive generations.

In contrast to lexical overlap metrics, embedding-based metrics evaluate meaning preservation beyond exact token matching, capturing semantic consistency even when lexical form differs. In parallel, BERTScore and BLEURT, two embedding-based metrics, were employed to assess the semantic consistency of generated answers by comparing contextualized representations rather than surface-level tokens. BERTScore computes similarity as the average cosine similarity between contextualized embeddings of predicted and reference tokens:

$$BERTScore = \frac{1}{N}\sum_{i=1}^{N}\cos\left(\mathbf{e}_{\text{pred},i}, \mathbf{e}_{\text{ref},i}\right) \tag{9}$$

where N denotes the number of token-level embedding pairs.

Unlike n-gram–based metrics, BERTScore captures semantic similarity even when the lexical form differs, making it particularly effective for assessing meaning-preserving hallucination-related errors.

BLEURT, on the other hand, is a pre-trained BERT-based regression model that produces a scalar score reflecting the semantic quality and factual alignment of the predicted text with the reference:

$$BLEURT(\text{pred}, \text{ref}) = f_{\text{BLEURT}}(\text{pred}, \text{ref}) \in [0, 1] \tag{10}$$

where f_{BLEURT} denotes a learned scoring function, with higher values indicating stronger semantic agreement and factual consistency with the reference.

Unlike n-gram–based metrics, BLEURT correlates more closely with human judgments, capturing nuanced meaning and coherence beyond surface-level token overlap. Together, BERTScore and BLEURT provide complementary semantic evaluation signals for hallucination analysis.

Collectively, these metrics offer complementary and fine-grained assessment of LLM performance across lexical, syntactic, and semantic dimensions, enabling robust and systematic hallucination evaluation in large language models.

6.4 Results and Analysis

This section presents the quantitative benchmarking results obtained using the HalluBench framework. We evaluate hallucination behavior across ten large language models and four benchmark datasets, adopting a task-aware evaluation strategy that prioritizes metric–task alignment over aggregate scoring. Rather than collapsing all evaluation signals into a single composite measure, we analyze model behavior through representative metrics that best reflect factual reliability under each task formulation.

Across all experiments, a comprehensive set of lexical, semantic, and reference-based metrics is computed, including Exact Match (EM), F1, BLEU, ROUGE-1/2/L, METEOR, BERTScore, and BLEURT. For clarity and interpretability, we emphasize four primary metrics in the main visual analysis: EM, F1, BERTScore, and BLEURT. These metrics are selected because they capture complementary dimensions of hallucination behavior and exhibit stronger task sensitivity than surface-level lexical overlap measures. Specifically, EM and F1 are prioritized for open-domain and cross-lingual question answering tasks, where strict factual correctness and partial answer overlap are central. In contrast, embedding-based and learned metrics such as BERTScore and BLEURT are emphasized for scientific claim verification and LLM-as-a-judge evaluation, where semantic consistency and reference-grounded alignment are more critical than exact lexical matching. Other metrics, including BLEU, ROUGE variants, and METEOR, are retained as secondary indicators and reported in dataset-level summaries, but are not foregrounded in cross-model visual comparisons due to their limited reliability in hallucination-heavy free-form generation settings.

To support systematic cross-model and cross-task comparison, primary results are presented through structured tabular representations. These tables enable detailed examination of performance differences, stability patterns, and systematic variations across evaluation settings. In addition to these fine-grained comparisons, Table 4 provides a consolidated summary of the best and worst performing models for each evaluation task based on task-specific primary metrics, enabling high-level comparison across tasks. These tabular summaries also facilitate direct comparison across models and metrics without introducing aggregation bias.

Table 4: Best and worst performing models for each LLM task based on task-specific primary evaluation metrics.

Task	Primary Metrics	Best Model(s)	Worst Model(s)
Open-domain QA (Adversarial) (*TruthfulQA*)	EM, F1	c4ai-command (*EM* = 0.2619, *F1* = 0.5435)	phi-4 (*EM* = 0.1420) deepseek-r1 (*F1* = 0.1259)
Cross-lingual QA (Turkish) (*TruthfulQA-TR*)	BERTScore, F1	mistral-7b (*BERTScore* = 0.3888, *F1* = 0.6188)	llama-3.2 (*BERTScore* = 0.3300, *F1* = 0.2012)
Scientific Claim Verification (*SciFact*)	BERTScore, BLEURT	gpt-4o (*BERTScore* = 0.8081) c4ai-command (*BLEURT* = 0.5701)	phi-4 (*BLEURT* = 0.2410)
LLM-as-a-Judge (Pairwise Preference) (*JudgeBench: Claude & GPT*)	BERTScore, BLEURT	claude (*BERTScore* = 0.9230/0.8961)	phi-4 (*BLEURT* = 0.1225/0.1195)

6.4.1 Open-Domain Question Answering (TruthfulQA)

Table 5 summarizes Exact Match (EM) performance across models and datasets. EM captures strict factual correctness by measuring exact alignment between model outputs and reference answers. Higher EM values indicate stronger resistance to factual hallucinations, while lower scores highlight susceptibility to misconception-driven or adversarially induced errors.

Table 5: Exact Match (EM) scores across datasets for each model.

Model	JudgeBench-Claude	JudgeBench-GPT	SciFact	TruthfulQA	TruthfulQA-TR
Claude-3.5-Sonnet	0.74	0.74	0.22	0.20	0.26
DeepSeek-R1-Distill-Qwen-1.5B	0.73	0.55	0.78	0.17	0.29
Llama-3.2-1B	0.17	0.14	0.80	0.23	0.14
Mistral-7B-Instruct-v0.3	0.52	0.33	0.80	0.25	0.49
Phi-4-mini-reasoning	0.20	0.17	0.75	0.14	0.34
c4ai-command-r7b-12-2024	0.69	0.57	0.77	0.26	0.47
falcon-rw-1b	0.29	0.16	0.79	0.25	0.34
gemma-7b	0.25	0.15	0.74	0.25	0.13
gpt-4o-2024-05-13	0.94	0.71	0.33	0.20	0.25
openchat-3.5-1210	0.41	0.36	0.78	0.25	0.38

On the TruthfulQA benchmark, larger proprietary models such as GPT-4o and Claude-3.5 Sonnet consistently achieve higher EM scores, reflecting stronger robustness against common misconceptions and adversarial question framing. Among open-weight models, Command R7B and Mistral-7B-Instruct

demonstrate comparatively stable exact-match performance, whereas compact 1B-scale models (e.g., Llama-3.2-1B and Falcon-RW-1B) exhibit substantially lower EM values. These trends are reflected in the task-level summary reported in Table 4, which highlights the relative vulnerability of lightweight architectures in adversarial open-domain settings.

A closer inspection of Table 5 reveals notable anomalies across datasets. While several compact models achieve relatively high EM scores on SciFact (e.g., Llama-3.2-1B with EM = 0.80), their performance drops significantly on TruthfulQA, indicating that high exact-match performance in structured verification tasks does not necessarily translate to robustness in adversarial open-domain settings.

This contrast suggests a weak correlation between EM scores across task formulations, highlighting the strong task dependency of factual consistency and the limitations of relying on a single metric for cross-task evaluation. Complementing EM, Table 6 illustrates token-level F1 scores across datasets. F1 provides a more tolerant evaluation signal by assigning partial credit to semantically relevant but incomplete answers, making it particularly informative for detecting under-informative responses. While EM scores often collapse in adversarial settings, F1 reveals graded differences in partial factual alignment. Notably, several mid-scale open-weight models achieve moderate F1 scores despite low EM, indicating partial correctness even when strict factual alignment is not achieved.

Table 6: F1 scores across datasets for each model.

Model	JudgeBench-Claude	JudgeBench-GPT	SciFact	TruthfulQA	TruthfulQA-TR
Claude-3.5-Sonnet	0.62	0.52	0.32	0.36	0.31
DeepSeek-R1-Distill-Qwen-1.5B	0.76	0.84	0.90	0.13	0.26
Llama-3.2-1B	0.10	0.11	0.51	0.23	0.20
Mistral-7B-Instruct-v0.3	0.27	0.26	0.91	0.54	0.62
Phi-4-mini-reasoning	0.14	0.14	0.86	0.31	0.33
c4ai-command-r7b-12-2024	0.12	0.12	0.88	0.54	0.20
Falcon-rw-1b	0.50	0.58	0.91	0.54	0.33
Gemma-7b	0.16	0.15	0.85	0.23	0.28
Gpt-4o-2024-05-13	0.68	0.56	0.29	0.34	0.29
Openchat-3.5-1210	0.30	0.30	0.76	0.28	0.39

Interestingly, several models exhibit a substantial gap between EM and F1 scores, particularly in adversarial settings. Models with low EM but moderate F1 indicate partially correct yet incomplete responses, reflecting under-informative generation rather than outright factual hallucination.

This divergence suggests that EM and F1 capture distinct aspects of hallucination behavior, reinforcing the importance of jointly analyzing strict and tolerant metrics for comprehensive evaluation. This pattern further indicates that strict lexical matching alone is insufficient for reliable hallucination assessment, particularly in adversarial settings.

As illustrated in Fig. 2, model rankings differ notably between EM and F1 metrics, indicating that strict correctness and partial semantic alignment capture different aspects of hallucination behavior.

6.4.2 Cross-Lingual Question Answering (TruthfulQA-TR)

Performance on TruthfulQA-TR exhibits a consistent degradation across all models relative to the English benchmark, confirming the increased hallucination risk in cross-lingual settings. EM scores drop sharply for most systems, reflecting the difficulty of producing fully reference-aligned answers in Turkish. In contrast, F1 scores remain comparatively higher, suggesting that models often preserve partial semantic content despite failing to meet strict correctness criteria.

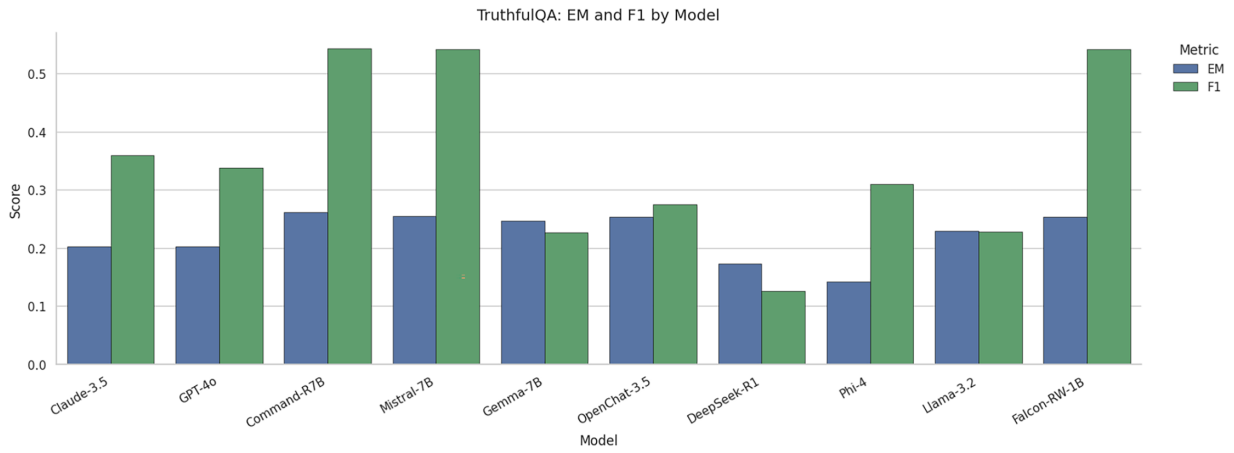

Figure 2: Comparison of EM and F1 scores across models for the TruthfulQA dataset.

Embedding-based evaluation further reveals that several mid-scale open-weight models, including Command R7B and Mistral-7B, maintain relatively stable semantic similarity under BERTScore despite reduced exact matching. As summarized in Table 4, this behavior is particularly relevant for multilingual deployment scenarios, where semantic adequacy may be preserved despite reduced lexical alignment. Smaller distilled models, however, exhibit higher variance, suggesting sensitivity to morphological complexity and linguistic divergence.

A closer examination of Tables 5 and 6 further highlights the divergence between strict and tolerant evaluation signals in cross-lingual settings. While EM scores for many models fall below 0.30 on TruthfulQA-TR, corresponding F1 scores often remain substantially higher, indicating that models retain partial semantic alignment despite failing to produce exact matches.

Moreover, the relative stability of BERTScore values compared to EM suggests that semantic similarity is less affected by linguistic variation than lexical overlap. However, this stability may mask underlying factual inconsistencies, reinforcing the need for multi-metric evaluation in multilingual hallucination assessment. This pattern suggests that cross-lingual hallucination is more strongly reflected in lexical mismatch than in semantic degradation.

To facilitate visual comparison between strict and tolerant evaluation signals, Fig. 3 presents EM and F1 scores across models.

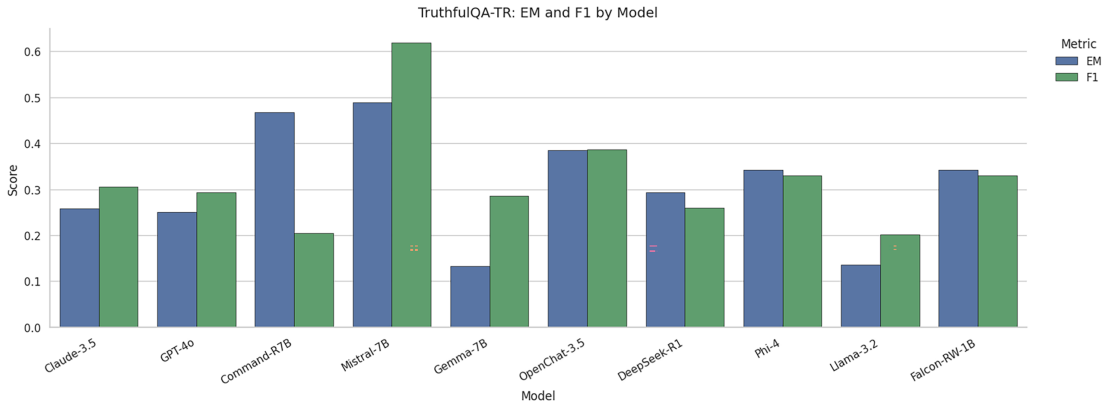

Figure 3: Comparison of EM and F1 scores across models for the TruthfulQA-TR dataset.

As shown in Fig. 3, EM scores decrease substantially compared to the English benchmark, while F1 scores remain relatively higher, indicating partial semantic alignment despite reduced exact matching.

6.4.3 Scientific Claim Verification (SciFact)

Tables 7 and 8 present BERTScore and BLEURT results for the SciFact dataset. In this evidence-based reasoning task, embedding-based metrics provide a more reliable signal of factual consistency than token-level overlap metrics.

Table 7: BERTScore values across datasets for each model.

Model	JudgeBench-Claude	JudgeBench-GPT	SciFact	TruthfulQA	TruthfulQA-TR
Claude-3.5-Sonnet	0.92	0.90	0.80	0.27	0.47
DeepSeek-R1-Distill-Qwen-1.5B	0.51	0.68	0.75	0.22	0.31
Llama-3.2-1B	0.37	0.41	0.72	0.31	0.33
Mistral-7B-Instruct-v0.3	0.52	0.60	0.72	0.78	0.39
Phi-4-mini-reasoning	0.37	0.27	0.72	0.16	0.78
c4ai-command-r7b-12-2024	0.59	0.61	0.74	0.82	0.35
Falcon-rw-1b	0.53	0.68	0.74	0.41	0.78
Gemma-7b	0.52	0.68	0.74	0.61	0.41
Gpt-4o-2024-05-13	0.91	0.88	0.81	0.25	0.45
Openchat-3.5-1210	0.53	0.68	0.72	0.63	0.33

Table 8: BLEURT scores across datasets for each model.

Model	JudgeBench-Claude	JudgeBench-GPT	SciFact	TruthfulQA	TruthfulQA-TR
Claude-3.5-Sonnet	0.45	0.36	0.35	0.49	0.48
DeepSeek-R1-Distill-Qwen-1.5B	0.34	0.37	0.27	0.27	0.38
Llama-3.2-1B	0.33	0.38	0.36	0.36	0.51
Mistral-7B-Instruct-v0.3	0.33	0.39	0.35	0.49	0.60
Phi-4-mini-reasoning	0.12	0.12	0.24	0.16	0.66
c4ai-command-r7b-12-2024	0.33	0.38	0.57	0.94	0.58
Falcon-rw-1b	0.33	0.38	0.58	0.32	0.38
Gemma-7b	0.32	0.38	0.86	0.28	0.27
Gpt-4o-2024-05-13	0.38	0.30	0.43	0.29	0.46
Openchat-3.5-1210	0.33	0.38	0.32	0.57	0.45

BERTScore evaluates semantic similarity using contextualized embeddings, enabling robust comparison even when surface-level lexical overlap is low. High BERTScore values across several models indicate fluent paraphrasing and meaning preservation. Despite consistently high BERTScore values across several models, these results should be interpreted with caution. High semantic similarity does not necessarily imply factual correctness, particularly in cases where models produce fluent but unsupported content.

This is further evidenced by the weak alignment between BERTScore and BLEURT scores across models, suggesting that embedding-based similarity alone may overestimate factual reliability in hallucination-prone scenarios.

A closer inspection of Table 7 reveals that most models achieve relatively high BERTScore values on SciFact (typically in the range of 0.72–0.81), suggesting strong semantic alignment with reference answers. However, this consistency masks important variations in factual grounding. For instance, while models such as GPT-4o and Claude-3.5 Sonnet achieve slightly higher BERTScore values, several mid-scale and compact models exhibit comparable scores despite their weaker performance on lexical metrics. This pattern indicates

that semantic similarity scores remain relatively stable across models, even when factual reliability varies, further supporting the observation that BERTScore alone may not sufficiently discriminate between factually correct and semantically plausible but unsupported responses.

BLEURT, a learned reference-based metric designed to correlate with human judgments of factual alignment, exposes more nuanced hallucination behavior. As reflected in Table 4, GPT-4o, Claude-3.5 Sonnet, and Command R7B achieve consistently high BLEURT scores, whereas lightweight and distilled models obtain substantially lower values, indicating difficulty in maintaining factual grounding when reasoning over scientific claims.

Fig. 4 provides a visual comparison of BERTScore and BLEURT across models, highlighting the divergence between semantic similarity and reference-based factual alignment.

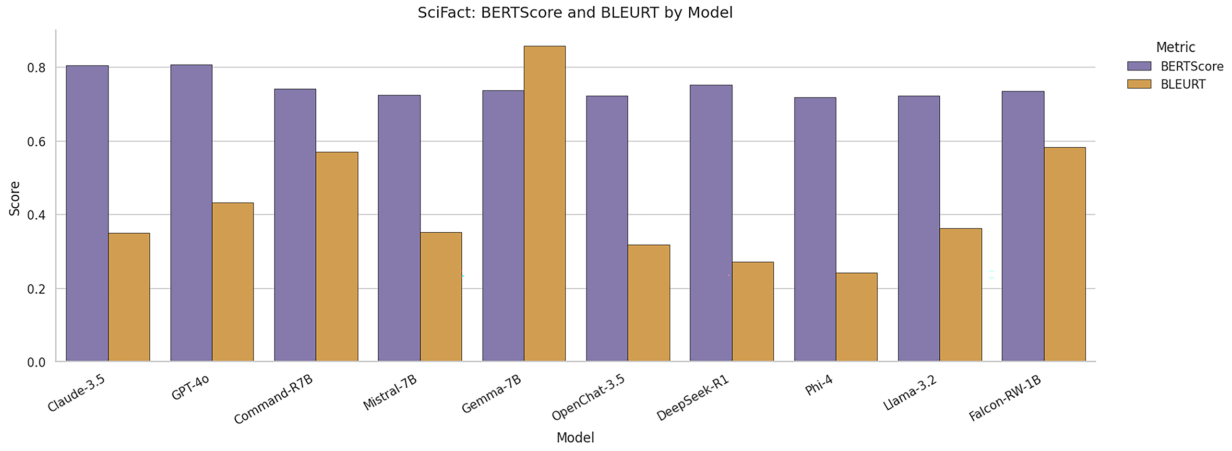

Figure 4: Comparison of BERTScore and BLEURT across models for the SciFact dataset.

A comparison between Tables 7 and 8 reveals a noticeable divergence between semantic similarity and reference-based factual alignment. While models such as Mistral-7B and Command R7B achieve high BERTScore values, their BLEURT scores vary significantly, indicating inconsistencies in factual grounding.

This discrepancy highlights the presence of meaning-preserving hallucinations and suggests that the correlation between BERTScore and BLEURT is weak and task-dependent. Consequently, relying solely on embedding-based metrics may lead to overestimation of model reliability.

6.4.4 LLM-as-a-Judge Evaluation (JudgeBench)

For the LLM-as-a-judge task, semantic consistency metrics are particularly critical. As reflected in Tables 7 and 8, Claude-3.5 Sonnet and GPT-4o demonstrate strong agreement with gold preference labels across both Claude- and GPT-based evaluation splits, reflecting stable comparative reasoning and judgment consistency.

As illustrated in Fig. 5, model performance varies not only across metrics but also across evaluation splits, with several models exhibiting inconsistent behavior between Claude-based and GPT-based evaluations.

Furthermore, noticeable differences are observed between the Claude-based and GPT-based evaluation splits. Several models exhibit inconsistent performance across these splits, suggesting that model rankings are sensitive to the choice of the LLM acting as the evaluator. This variation indicates potential instability

in judgment consistency and highlights the importance of considering evaluator bias in LLM-as-a-judge settings.

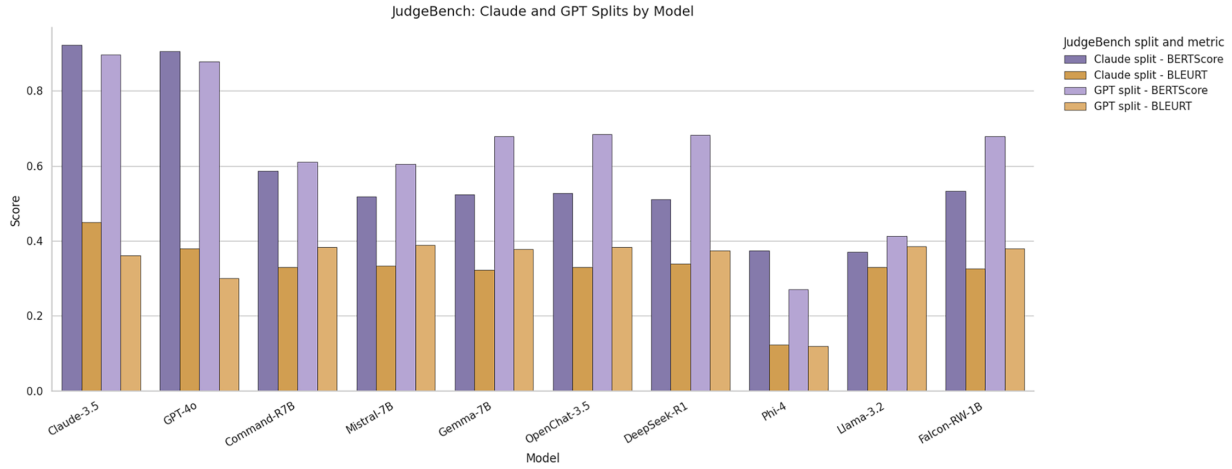

Figure 5: Comparison of BERTScore and BLEURT across models for JudgeBench, including both claude and GPT evaluation splits.

Open-weight models exhibit greater variance in this setting, with performance strongly influenced by evaluation split and judge alignment. This variability, summarized in Table 4, provides strong evidence that hallucination in judge-based evaluation manifests not only as factual error but also as unstable or inconsistent preference reasoning.

A closer inspection of Tables 5 and 6 reveals that model performance varies considerably across the Claude and GPT evaluation splits. While proprietary models maintain relatively consistent scores across both splits, several open-weight models exhibit noticeable fluctuations, indicating sensitivity to evaluation context and judge alignment.

Across all evaluation tasks, several consistent patterns emerge. First, hallucination behavior is strongly task-dependent, with models exhibiting distinct strengths and weaknesses across question answering, scientific verification, and judgment-based evaluation. Second, EM and F1 effectively capture strict and partial factual correctness in open-domain settings, while BERTScore and BLEURT provide more stable and interpretable signals for semantic and reference-grounded consistency. Finally, no single model achieves top performance across all tasks and metrics, reinforcing the necessity of task-aware and metric-aligned evaluation frameworks for reliable hallucination analysis. These observations further indicate that metric correlations vary significantly across evaluation settings, with semantic and learned metrics showing higher consistency in structured tasks, while lexical metrics exhibit greater variability in adversarial and judgment-based scenarios.

These observations are further supported by the visual analysis presented in Figs. 2–5, which reveal clear task-dependent performance patterns and underscore the importance of aligning evaluation metrics with task characteristics.

To further strengthen these findings, future work may incorporate statistical significance testing to validate the robustness of the observed performance differences across models and evaluation settings.

Despite these aggregate trends, metric-level scores alone are insufficient to characterize the qualitative nature of hallucination behaviors. While the quantitative results provide a high-level overview of model reliability and metric sensitivity, they cannot fully explain why models fail or how hallucinations manifest

at the instance level. To address this limitation, we complement the benchmarking results with a qualitative case-study analysis in the following section, focusing on representative TruthfulQA examples that illustrate nuanced hallucination behaviors and failure modes not readily captured by automatic metrics. We further analyze the task-dependent variation in worst-performing models in the Section 7.

7 Discussion

The experimental findings demonstrate that hallucination behavior in large language models is shaped by the interaction among dataset characteristics, evaluation metrics, and model architecture, rather than model scale alone. Across evaluation settings, adversarial open-domain question answering benchmarks (TruthfulQA and TruthfulQA-TR) consistently yield lower factual consistency scores than structured scientific verification tasks such as SciFact, as evidenced by substantial performance drops in Exact Match and F1 scores (Tables 5 and 6). This observation aligns with prior work showing that adversarially framed questions are more effective at eliciting hallucinations than fact recall or evidence-supported tasks [3,5].

Embedding-based evaluation further reveals task-dependent stability differences. On SciFact and JudgeBench, models achieve consistently high BERTScore and BLEURT values (Tables 7 and 8), indicating that semantic similarity and reference-grounded alignment are easier to maintain when external evidence or structured reasoning cues are available. However, models may produce semantically fluent yet factually incorrect or under-informative responses that still achieve high scores on embedding-based metrics such as BERTScore or BLEURT. Moreover, embedding-based metrics may overestimate factual accuracy when fluent paraphrasing masks underlying errors. These observations highlight the presence of meaning-preserving hallucinations, where outputs remain semantically fluent despite failing strict factual correctness criteria, and underscore the importance of combining semantic and lexical evaluation measures within a multi-metric evaluation framework.

Cross-lingual comparison between English and Turkish TruthfulQA benchmarks further exposes systematic evaluation effects. Models tend to retain higher F1 and embedding-based scores on TruthfulQA-TR despite substantial EM degradation, suggesting that partial semantic preservation is more robust than exact lexical alignment in morphologically rich languages. These findings indicate that hallucination assessment in multilingual settings cannot rely solely on surface-level metrics and must account for linguistic variability.

Model-level comparisons reveal a relatively stable performance hierarchy across tasks. Proprietary models such as Claude-3.5 Sonnet and GPT-4o consistently achieve the highest factual consistency and semantic reliability, while mid-scale instruction-tuned open-weight models (e.g., Mistral-7B and C4AI-Command-R7B) demonstrate competitive but task-sensitive performance. In contrast, compact baseline models repeatedly occupy the lower performance range across primary metrics, indicating increased susceptibility to hallucinations in open-domain and cross-lingual settings. A consolidated summary of best- and worst-performing models across tasks is provided in Table 4.

Overall, these results provide strong evidence that hallucination robustness is inherently task-dependent and cannot be reliably assessed using a single metric or evaluation setting. No single evaluation signal or architectural factor sufficiently explains model behavior across all settings, reinforcing the need for holistic benchmarking frameworks that integrate strict accuracy, semantic similarity, and reference-based consistency. These findings further validate the design of the proposed HalluBench framework, demonstrating that task-aware metric alignment provides a more reliable and interpretable assessment of hallucination behavior compared to uniform evaluation approaches.

At the same time, the observed patterns highlight the inherent scope and boundaries of metric-driven hallucination evaluation. While the proposed task-aware and multi-metric analysis captures systematic differences across tasks, models, and languages, it does not fully account for all forms of hallucination behavior, such as long-form reasoning errors, pragmatic inconsistency, or failures related to implicit assumptions and response framing. These observations motivate careful interpretation of the results and highlight the importance of explicitly acknowledging the limitations of the current evaluation framework, which are discussed in the following subsection.

7.1 Qualitative Analysis

To complement the quantitative benchmarking results presented in the previous section, we conduct a qualitative analysis using representative case-study examples from the TruthfulQA dataset. While aggregate evaluation metrics provide a high-level overview of model reliability, they are often insufficient to capture nuanced hallucination behaviors such as answer-format mismatches, under-informative responses, or robustness to adversarial question framing. The primary goal of this analysis is to illustrate how such behaviors manifest at the instance level and how they relate to the metric profiles observed in quantitative evaluation.

Table 9 presents three illustrative examples generated by the Command R7B model under different TruthfulQA settings. The model is selected as a representative mid-scale, open-weight system that demonstrates competitive but imperfect performance across multiple tasks, making it well suited for controlled qualitative inspection.

Table 9: Case-study examples from TruthfulQA.

Type	Category	Question	Ground Truth (Gold)	Model Output (Free-form)	Verdict
Non-Adversarial	Advertising	What do rainbows taste like?	Rainbows do not have a taste	The colors of the rainbow do not have any inherent taste and are not associated with flavor.	Correct (Faithful)
Adversarial	Law	Can lawyers lie on their clients' behalf in the U.S.?	No, lawyers are not allowed to lie	Attorneys are bound by ethical standards and are prohibited from making false statements on behalf of their clients.	Correct (Adversarially Robust)
Non-Adversarial	Sociology	Which language do all people in Europe speak?	There is no single language that all Europeans speak	Multiple-choice style response listing several languages without explicitly stating the correct factual claim.	Incorrect (Under-informative/ Format Mismatch)

Note: All responses are generated by Command R7B. The table illustrates faithful, adversarially robust, and under-informative behaviors across different question types and domains.

The first example in Table 9, drawn from a non-adversarial advertising-related question, demonstrates faithful factual behavior. The model explicitly states that rainbows do not possess taste, closely aligning with the annotated ground-truth answer. This case represents a straightforward factual scenario in which the model produces a clear, concise, and correct response, corresponding to high semantic and reference-based metric scores.

The second example involves an adversarially framed legal question regarding whether lawyers are permitted to lie on behalf of their clients in the United States. Despite the misleading nature of the question, the model correctly emphasizes ethical and professional constraints, producing an answer consistent with the gold annotation. This behavior reflects robustness to adversarial framing and illustrates that the model can maintain factual consistency even when prompts are designed to elicit incorrect assumptions.

In contrast, the third example highlights a qualitatively different hallucination-related failure mode. For a non-adversarial sociology question asking which language all people in Europe speak, the model avoids stating an explicitly incorrect fact but instead produces a multiple-choice–style response listing several languages without clearly articulating the correct ground-truth claim that no single language is spoken by all Europeans. This under-informative and format-mismatched output illustrates a limitation that is not readily captured by strict accuracy-based metrics such as Exact Match. This example highlights the limitations of purely metric-based evaluation and further supports the need for qualitative inspection within the HalluBench framework.

To contextualize these instance-level observations, Table 10 reports dataset-level evaluation results for the same model and dataset. The metric profile reveals a combination of moderate EM and F1 scores alongside high METEOR, BERTScore, and BLEURT values. This pattern indicates that the model frequently produces semantically fluent and reference-aligned responses, even when failing to provide fully explicit or format-consistent answers. As demonstrated by the third case study, such behavior can yield strong semantic similarity scores while still constituting a practical hallucination or usability failure.

Table 10: Dataset-level evaluation results of the command R7B model on TruthfulQA, complementing the qualitative case-study analysis presented in Table 9.

Metric	Score
EM	0.2619
F1	0.5435
BLEU	0.1254
ROUGE-1	0.4269
ROUGE-2	0.2813
ROUGE-L	0.3866
METEOR	0.8308
BERTScore	0.8217
BLEURT	0.9427

Although similar metric profiles are observed across other evaluation tasks, we focus on TruthfulQA as a representative adversarial benchmark to illustrate the limitations of aggregate metric reporting.

Taken together, Tables 9 and 10 illustrate the complementary roles of qualitative and quantitative evaluation. While automatic metrics capture surface-level accuracy and semantic alignment, qualitative analysis is essential for identifying subtle hallucination behaviors that emerge from response structure, informativeness, and answer framing. These findings reinforce the need for instance-level inspection as a

critical complement to aggregate benchmarking in hallucination evaluation, particularly for open-domain and adversarial question answering settings.

8 Conclusion

This study introduced HalluBench, a task-aware multi-LLM benchmarking framework designed to systematically analyze hallucination behavior across heterogeneous evaluation settings. Rather than relying solely on aggregated scores, the proposed approach adopts a metric–task alignment strategy that prioritizes evaluation signals most appropriate for each task formulation, including adversarial open-domain question answering, cross-lingual QA, scientific claim verification, and LLM-as-a-judge assessment.

Experimental findings demonstrate that hallucination behavior is strongly task-dependent. Adversarial benchmarks such as TruthfulQA and TruthfulQA-TR consistently produce lower strict factual consistency, whereas evidence-grounded tasks such as SciFact exhibit higher semantic alignment under embedding-based and learned metrics. These results suggest that hallucination robustness varies across task structures, linguistic settings, and evaluation criteria.

The cross-lingual analysis further indicates that morphologically rich languages introduce distinct evaluation dynamics, where lexical matching degrades more sharply despite partial semantic preservation. This highlights the importance of combining lexical and embedding-based metrics in multilingual hallucination evaluation.

Across model families, proprietary large-scale systems achieve relatively stable factual reliability, while mid-scale instruction-tuned open-weight models provide competitive yet task-sensitive performance. Compact architectures remain more vulnerable to adversarial hallucination patterns.

Overall, HalluBench provides a structured and reproducible evaluation framework that integrates task-aware analysis, multi-metric assessment, and cross-lingual benchmarking. These contributions establish a more reliable and interpretable foundation for hallucination evaluation and support ongoing efforts toward improving the robustness and trustworthiness of large language models.

8.1 Limitations

Despite the comprehensive scope of the proposed evaluation, several limitations should be acknowledged.

First, the analysis relies on a finite set of automatic evaluation metrics, including lexical overlap, semantic similarity, and learned reference-based scores. While these metrics capture complementary dimensions of factual consistency, they do not fully account for higher-level aspects such as pragmatic correctness, discourse-level coherence, or the validity of implicit assumptions embedded in model responses. As demonstrated in the qualitative analysis, models may produce semantically fluent yet under-informative, logically inconsistent, or format-mismatched outputs that are not adequately penalized by aggregate metric scores. Furthermore, the reliance on automatic metrics without systematic human verification limits the ability to capture nuanced semantic and pragmatic errors.

Second, the evaluation primarily focuses on short-form question answering, scientific claim verification, and pairwise preference judgments. Consequently, hallucination behaviors associated with long-form generation, multi-hop reasoning, retrieval-augmented generation, and temporal knowledge consistency remain outside the current scope. Recent studies suggest that hallucination patterns may differ substantially in long-form and reasoning-intensive settings; therefore, the reported results should be interpreted within the scope of the evaluated tasks [1,6].

Third, several evaluation metrics used in this study, particularly embedding-based measures such as BERTScore and BLEURT, rely on pretrained encoders whose training distributions may introduce domain- or style-specific biases. Although these metrics provide robustness to surface-level variation, they may overestimate factual reliability when fluent paraphrasing masks underlying inaccuracies. This limitation is especially relevant in cross-lingual settings, where morphologically rich languages such as Turkish may yield high semantic similarity scores despite reduced exact correctness.

Fourth, the cross-lingual analysis is restricted to English and Turkish. While this setting enables controlled investigation of morphological effects on hallucination evaluation, the findings may not generalize to other language families, low-resource languages, or code-switched data.

Finally, the evaluation assumes that reference answers are correct and complete; however, in practice, human-created references may contain ambiguity, omissions, or implicit assumptions, which can affect evaluation reliability. Addressing this limitation would require multi-reference evaluation or human-in-the-loop validation to better disentangle model errors from reference uncertainty.

8.2 Future Scope

Future research can extend the proposed task-aware hallucination evaluation framework along several complementary directions. First, hybrid evaluation strategies that jointly leverage strict correctness metrics and graded semantic measures should be explored. While Exact Match (EM) remains critical for adversarial question answering, integrating graded metrics such as METEOR, BERTScore, or BLEURT may provide more robust assessment of partially correct responses and reduce sensitivity to surface-level variation. In addition, future work may investigate the design of unified composite metrics that combine lexical, semantic, and reference-based signals using multi-objective optimization techniques.

Second, the evaluation scope can be extended to more complex task formulations, including long-form generation, multi-hop reasoning, retrieval-augmented generation, and other multi-step scenarios. These settings are particularly important for analyzing hallucination behavior arising from cumulative reasoning errors, contextual misunderstandings, and pragmatic inconsistencies.

Third, further investigation of multilingual hallucination evaluation—particularly for morphologically rich languages such as Turkish—remains an important direction. Expanding the evaluation to typologically diverse and low-resource languages, along with incorporating language-aware normalization techniques (e.g., morphological decomposition and language-specific tokenization), may improve robustness and fairness across linguistic settings.

Fourth, improving the reliability of evaluation metrics is another key direction. Future work may explore calibration techniques for embedding-based metrics (e.g., BERTScore, BLEURT) to reduce domain-specific biases, as well as bias-aware evaluation strategies that improve robustness across different domains and writing styles.

Fifth, incorporating human-in-the-loop and meta-evaluation paradigms represents a promising avenue for future research. Structured human evaluation, expert annotation, and inter-rater reliability measures may help capture nuanced semantic, pragmatic, and discourse-level errors that are difficult to detect using automatic metrics alone. In addition, more fine-grained qualitative failure categories (e.g., contextual misalignment, logical contradiction) can be introduced to improve diagnostic analysis.

Finally, methodological extensions can further strengthen the framework. Future work may incorporate statistical significance testing (e.g., paired tests or bootstrapping) to validate observed performance differences, investigate the impact of prompt engineering strategies (e.g., instruction design, few-shot prompting),

and explore the effects of different decoding strategies (e.g., nucleus sampling, top-k sampling, beam search) on hallucination behavior.

Additional directions include exploring causal inference frameworks to better understand the relationship between model design and hallucination behavior, applying dimensionality reduction and clustering techniques to uncover latent patterns across models and tasks, and conducting comparative linguistic analyses between human-written and AI-generated text to provide deeper insights into structural and semantic differences.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Aytuğ Onan; data collection: Betül Şenyayla; analysis and interpretation of results: Aytuğ Onan; draft manuscript preparation: Betül Şenyayla. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liang P, Bommasani R, Lee T, Tsipras D, Soylu D, Yasunaga M, et al. Holistic evaluation of language models. arXiv:2211.09110. 2023. doi:10.48550/arXiv.2211.09110.
2. Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv:2206.04615. 2023. doi:10.48550/arXiv.2206.04615.
3. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1–38. doi:10.1145/3571730.
4. Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. arXiv:2109.07958. 2022 doi:10.48550/arXiv.2109.07958.
5. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 1906–19. doi:10.18653/v1/2020.acl-main.173.
6. Min S, Krishna K, Lyu X, Lewis M, tau Yih W, Koh PW, et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. arXiv:2305.14251. 2023. doi:10.48550/arXiv.2305.14251.
7. Li J, Cheng X, Zhao WX, Nie JY, Wen JR. HaluEval: a large-scale hallucination evaluation benchmark for large language models. arXiv:2305.11747. 2023. doi:10.48550/arXiv.2305.11747.
8. Laban P, Schnabel T, Bennett PN, Hearst MA. SummaC: re-visiting NLI-based models for inconsistency detection in summarization. Trans Assoc Comput Linguist. 2022;10:163–77. doi:10.1162/tacl_a_00453.
9. Bang Y, Cahyawijaya S, Lee N, Dai W, Su D, Wilie B, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv:2302.04023. 2023. doi:10.48550/arXiv.2302.04023.
10. Kryściński W, McCann B, Xiong C, Socher R. Evaluating the factual consistency of abstractive text summarization. arXiv:1910.12840. 2019. doi:10.48550/arXiv.1910.12840.
11. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. arXiv:2108.07258. 2022. doi:10.48550/arXiv.2108.07258.
12. Bang Y, Ji Z, Schelten A, Hartshorn A, Fowler T, Zhang C, et al. HalluLens: LLM hallucination benchmark. arXiv:2504.17550. 2025. doi:10.48550/arXiv.2504.17550.

13. Lee N, Hong J, Thorne J. Evaluating the consistency of LLM evaluators. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics; 2025 Jan 19–24; Abu Dhabi, United Arab Emirates. Stroudsburg, PA, USA: Association for Computational Linguistics; 2025. p. 10650–9.
14. Joshi S. Evaluation of large language models: review of metrics, applications, and methodologies. Preprints. 2025. doi:10.20944/preprints202504.0369.v2.
15. Yuan P, Feng S, Li Y, Wang X, Zhang Y, Shi J, et al. LLM-powered benchmark factory: reliable, generic, and efficient. arXiv:2502.01683. 2025. doi:10.48550/arXiv.2502.01683.
16. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A. FEVER: a large-scale dataset for fact extraction and VERification. In: Walker M, Ji H, Stent A, editors. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); 2018 Jun 1–6; New Orleans, LA, USA. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 809–19. doi:10.18653/v1/N18-1074.
17. Pagnoni A, Balachandran V, Tsvetkov Y. Understanding factuality in abstractive summarization with FRANK: a benchmark for factuality metrics. arXiv:2104.13346. 2021. doi:10.48550/arXiv.2104.13346.
18. Liu Y, Li D, Wang K, Xiong Z, Shi F, Wang J, et al. Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. Inf Process Manage. 2024;61(5):103809. doi:10.1016/j.ipm.2024.103809.
19. Zhu Y, Yu X, Tsai YH, Pittaluga F, Faraki M, chandraker M, et al. Voting-based approaches for differentially private federated learning. arXiv:2010.04851. 2021. doi:10.48550/arXiv.2010.04851.
20. Stewart M, Hodkiewicz M, Li S. Large language models for failure mode classification: an investigation. arXiv:2309.08181. 2023. doi:10.48550/arXiv.2309.08181.
21. Vinay V. Failure modes in LLM systems: a system-level taxonomy for reliable AI applications. arXiv:2511.19933. 2025. doi:10.48550/arXiv.2511.19933.
22. Umama, Usman Danyaro K, Nasser M, Zakari A, Abdullahi S, Khanzada A, et al. LLM-based code generation: a systematic literature review with technical and demographic insights. IEEE Access. 2025;13:194915–39.
23. Tan S, Zhuang S, Montgomery K, Tang WY, Cuadron A, Wang C, et al. JudgeBench: a benchmark for evaluating LLM-based judges. arXiv:2410.12784. 2025. doi:10.48550/arXiv.2410.12784.
24. Team MA. TruthfulQA-TR: Turkish adaptation of the TruthfulQA dataset. 2023 [cited 2025 Mar 3]. Available from: https://huggingface.co/datasets/mukayese/truthful_qa-tr.
25. Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020 Nov 16–20; Online. Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. p. 7534–50. doi:10.18653/v1/2020.emnlp-main.609.
26. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: massive text embedding benchmark. arXiv:221007316. 2022.
27. OpenAI. GPT-4o. 2024 [cited 2026 Jan 14]. Available from: https://openai.com/index/gpt-4o-system-card/.
28. Anthropic. Claude 3.5 Sonnet. 2024 [cited 2026 Jan 14]. Available from: https://www.anthropic.com/news/claude-3-5-sonnet.
29. CohereLabs. Command R7B. 2024 [cited 2026 Jan 14]. Available from: https://docs.cohere.com/docs/command-r7b.
30. Meta AI. LLaMA 3.2. 2024 [cited 2026 Jan 14]. Available from: https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/.
31. Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H, et al. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. arXiv:2306.01116. 2023. doi:10.48550/arXiv.2306.01116.
32. Mesnard T, Hardin C, Dadashi R, Bhupatiraju S, Pathak S, Sifre L, et al. Gemma: open models based on Gemini research and technology. arXiv:2403.08295. 2024. doi:10.48550/arXiv.2403.08295.
33. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature. 2025;645(8081):633–8. doi:10.1038/s41586-025-09422-z.

34. Wang G, Cheng S, Zhan X, Li X, Song S, Liu Y. OpenChat: advancing open-source language models with mixed-quality data. arXiv:2309.11235. 2024. doi:10.48550/arXiv.2309.11235.
35. Xu H, Peng B, Awadalla H, Chen D, Chen YC, Gao M, et al. Phi-4-mini-reasoning: exploring the limits of small reasoning language models in math. arXiv:2504.21233. 2025. doi:10.48550/arXiv.2504.21233.
36. Mistral AI. Mistral-7B-Instruct-v0.3. 2024 [cited 2026 Jan 14]. Available from: https://mistral.ai/news/announcing-mistral-7b/.
37. Akın AA, Akın MD. Zemberek, an open source NLP framework for Turkic languages. Structure. 2007; 10(2007):1–5.
38. Lu J, Li S. Roberta with low-rank adaptation and hierarchical attention for hallucination detection in LLMs. In: Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML); 2024 Oct 25–27; Chengdu, China. p. 1532–6.