

Measuring Hallucination Rate in LLMs Across Domains and Model Versions

Nikita B. Emberi, V V S Aakash Kotha

Abstract

Hallucination—where a language model produces fluent but factually incorrect content—poses a critical challenge for reliable AI deployment. This paper presents a systematic study of hallucination rates across multiple knowledge domains and large language model (LLM) versions. We collected a new evaluation dataset of 400 fact-based questions spanning five domains (general knowledge, science, history, pop culture, and politics) and obtained 1,200 answers from three state-of-the-art LLMs (OpenAI GPT-3.5, OpenAI GPT-4, and Anthropic Claude). Each response was manually annotated for factual correctness (binary label and a graded “FactScore” for partial accuracy), and linguistic features such as citation-style phrases and hedging were recorded. Our analysis reveals that hallucination frequency varies significantly by domain and by model. In particular, abstract domains (e.g. politics, history) exhibit higher hallucination rates than concrete domains like science, and the newest model (GPT-4) hallucinated substantially less frequently than the older GPT-3.5 model. Quantitatively, GPT-4’s hallucination rate was less than half that of GPT-3.5 in our experiments, confirming improved factuality in newer model versions. Conversely, question complexity (measured by input length) showed no strong correlation with hallucination likelihood after controlling for domain and question type. These findings underscore the importance of evaluating LLM outputs across diverse scenarios. By identifying where and when hallucinations occur, our work informs development of safer, more trustworthy AI systems. **GitHub Repository Link:** <https://github.com/AakashKotha/llm-hallucination-research>

1. Introduction

Large Language Models (LLMs) like GPT-4 and Claude have achieved remarkable fluency, but they often hallucinate—that is, generate confident-sounding statements that are factually incorrect or ungrounded [1]. Such hallucinations can range from minor errors (e.g. wrong dates or names) to wholly fabricated answers, undermining the reliability of LLMs in applications such as education, science, and healthcare. For example, a recent study found that ChatGPT’s GPT-3.5 model produced a substantial fraction of non-existent or incorrect references when asked to generate scholarly citations

[2]. Hallucinated outputs are problematic because they can mislead users, especially when delivered with the model’s characteristic fluency and apparent confidence. As LLMs become integrated into high-stakes domains, understanding and mitigating these failure modes is critical for AI safety and transparency.

Despite the clear risks, measuring and characterizing hallucinations remains challenging. There is no single agreed-upon metric for “factual accuracy” across all tasks, and prior research has mostly evaluated models on narrow benchmarks or via anecdotal examples. Many early studies emphasized overall task accuracy or focused on improving instruction-following, rather than directly analyzing hallucination phenomena. In the context of text generation, hallucination is typically defined as content that appears plausible but cannot be verified or is outright false [1]. Researchers distinguish between intrinsic hallucinations (where the output contradicts the provided source or prompt) and extrinsic hallucinations (where the output includes information that is not supported by any source) [1]. Hallucinations have been widely observed in tasks like abstractive summarization and open-domain question answering. For instance, abstractive summarization models on the XSum dataset were reported to produce unfaithful information in over 60% of summaries [3], and even instruction-tuned chat models frequently fabricate details in their responses.

Encouragingly, newer model generations have shown improvement in factuality on certain benchmarks. OpenAI’s GPT-4, for example, has demonstrated significantly lower hallucination rates than its predecessor GPT-3.5 on various evaluations[2]. Nevertheless, the problem is far from solved, and systematic, fine-grained evaluations are needed. Notably, few works have systematically studied how hallucination behavior varies across different knowledge domains or across successive model versions. Does a model that is reliable in one subject area (say, science questions) falter in another (history or politics)? And to what extent do the massive recent LLMs actually overcome the hallucination tendencies of earlier models? These questions have practical importance for deploying LLMs in domain-specific settings and for tracking progress in model development. In this work, we address these gaps by conducting a targeted evaluation of hallucination rates across multiple domains and model versions. We introduce a new dataset and evaluation protocol to measure when an LLM’s answer is factual versus a hallucination. Unlike prior studies that might use a single benchmark or

only one model, our study covers a breadth of scenarios: five distinct knowledge domains and three generational levels of LLMs. We also incorporate analysis of certain linguistic markers or response features (such as the presence of citation-style phrases, or expressions of certainty/uncertainty) to examine whether they correlate with factual accuracy. By analyzing when and why hallucinations occur, our aim is to inform the development of more robust, truth-consistent language models.

2. Motivation

Hallucination in LLMs is not just a theoretical nuisance—it poses tangible risks in real-world deployments. In domains like medicine, law, or journalism, an AI-generated false statement can have serious consequences if taken at face value. One major concern is that hallucinated outputs are often delivered in a deceptively believable manner; the model’s fluent language and confident tone can make it difficult for users to discern fact from fiction. This is exacerbated by the growing trust users place in AI assistants. If an educational tutor bot confidently teaches a historical falsehood, or a clinical assistant suggests a non-existent study as evidence, users may be misled with high confidence. As such, identifying and reducing hallucinations is crucial for responsible AI adoption.

From a research standpoint, understanding hallucinations can shed light on the limitations of current model architectures and training regimes. Hallucinations often indicate that the model is operating outside the bounds of its knowledge or has learned spurious patterns. By examining the conditions under which hallucinations occur (e.g., certain topics that the model was not adequately trained on, or question formulations that confuse the model), we gain insight into how these models store and retrieve factual information. This can guide improvements in training data curation or model fine-tuning to address factual consistency. Moreover, developing reliable hallucination detection methods could enable AI systems to know when they don’t know—an important step toward safer deployment. In summary, reducing hallucinations matters both for practical reliability (preventing the spread of misinformation, increasing user trust) and for scientific progress (illuminating model weaknesses and helping to fix them).

Another motivation for our study is the gap in current research regarding cross-domain and cross-model comparisons of hallucination behavior. Most existing evaluations of factual accuracy have looked at a single model on a single task or benchmark. For example, a model might be tested on a medical Q&A dataset or a trivia quiz dataset in isolation. However, this leaves open questions: does a model that hallucinates on open-ended trivia also hallucinate on scientific questions at the same rate? Are certain domains inherently harder for LLMs to be truthful about (perhaps due to ambiguity or lack of training data in those areas)? Likewise, while it is generally assumed that newer models

like GPT-4 are more factual than older ones, the extent of improvement and its consistency across different conditions is not well quantified. Our work is motivated by these unknowns. By rigorously comparing domain-specific hallucination rates and examining multiple model versions (released at different times), we aim to provide a clearer picture of where we stand in mitigating hallucinations. This can inform researchers and practitioners about which domains or query types need extra caution (or additional model training) and whether upgrading to the latest model truly yields reliability gains in their use-case. Ultimately, the motivation is to pave the way for LLMs that can be trusted not to “make things up” regardless of the context in which they are used.

3. Background (Literature Survey)

The phenomenon of AI hallucinations has been recognized across various natural language generation tasks, and a growing body of literature seeks to define, detect, and reduce these errors. Hallucination in text generation was formally discussed in machine translation and summarization contexts as early as the late 2010s [4]. Ji et al. (2023) provide a comprehensive survey of hallucinations in NLG, classifying them and reviewing mitigation strategies [5]. In general, a hallucination refers to generated content that is unfaithful to the source or to reality – the model outputs information that is not supported by the input or by factual knowledge. This can manifest as intrinsic hallucinations (internally inconsistent with provided data) or extrinsic hallucinations (fabricated facts introduced). Both types degrade the utility of LLM responses, especially as outputs become longer and more complex.

A consistent finding in prior work is the prevalence of hallucinations in unconstrained generation tasks. Abstractive summarization systems, for instance, often include details not present in the source document. Maynez et al. (2020) reported that on the XSum news summary dataset, over 60% of model-generated summaries contained some content not supported by the article [6]. Even high-performing summarization models frequently distort or introduce facts, prompting a line of research into factual consistency metrics. Early approaches to evaluate summary factuality include automatic metrics like FactCC (Kryściński et al., 2019), which uses a learned entailment model to judge if a summary is entailed by the source, and QAGS (Wang et al., 2020), which uses question answering to probe summary correctness. These metrics signaled the community’s realization that standard NLG metrics (like ROUGE or BLEU) do not penalize factual errors.

In the realm of dialogue and open-domain question answering, the introduction of powerful LLMs has made the hallucination problem more prominent to end-users. ChatGPT and similar models sometimes produce entirely fictional explanations or sources, a behavior widely publicized when users found chatbots pro-

viding invented biographies or references. For example, academic and medical communities have noted that ChatGPT can generate fabricated references that look plausible but do not actually exist [2]. In one comparative study in the medical domain, ChatGPT (GPT-3.5) hallucinated references at a rate of about 40%, whereas GPT-4 reduced this to around 28% [2] – a substantial improvement but still far from acceptable for rigorous use. Another analysis by OpenAI on truthfulness benchmarks (TruthfulQA) showed that GPT-4 achieved much higher accuracy on adversarial truth-testing questions than GPT-3.5, highlighting progress in mitigating some forms of hallucination. Nonetheless, even GPT-4 is not immune to errors: it can be tricked by sufficiently novel or ambiguous queries, and remains constrained by its training cutoff and knowledge scope.

Related work on hallucination mitigation has explored various angles: improving training data (to reduce knowledge gaps and contradictions), incorporating retrieval (so the model has access to a knowledge source and is less likely to invent facts), and refining decoding strategies (to avoid the model’s propensity to fill in uncertain answers with its best guess). For instance, tuning models with human feedback on factuality, or post-editing responses to include verifiable evidence, have shown reductions in hallucination frequency. Another practical strategy in deployed systems is to have the model express uncertainty (e.g., by saying “I’m not sure” or refusing to answer) when it lacks confidence. Some LLMs like Claude are observed to be more conservative, often prefacing with caution if uncertain. However, balancing refusals and helpfulness is tricky – models that are too eager to abstain might avoid answering legitimate queries.

A key gap in the literature, and one that our work targets, is the systematic evaluation of hallucinations across domains and model versions. Most of the above studies analyze a single model (usually one of the GPT series or a comparable model) and often on a single domain or dataset. There has been relatively little published on comparing multiple LLMs side-by-side on the same hallucination evaluation, especially across different knowledge areas. One recent effort in this direction is the introduction of benchmarks like HaluEval and HalluLens (2024) which aim to benchmark hallucination propensity of various models on controlled tasks [7]. These benchmarks found, for example, that GPT-4 (sometimes referred to as “GPT-4o” in the literature for a specific tuned version) had the best factual accuracy among tested models, and that larger models generally hallucinate less [7]. They also noted that certain models (like some Anthropic Claude variants) tend to err on the side of caution by refusing to answer dubious queries, which can inflate precision scores at the cost of higher refusal rates [7]. Our work complements these findings by focusing on everyday fact-based questions across domains and explicitly measuring hallucination rates in a comparative framework. By surveying prior research, we underscore the need for our study: while it is known that hallucinations are a serious issue and

that newer models are making progress, we lack detailed understanding of which contexts are most prone to hallucinations and how big the improvement from model to model really is on a consistent set of tasks. The following sections describe our research questions and methodology to address these issues.

4. Data Description

To investigate the above questions, we constructed a dedicated evaluation dataset and annotation scheme that captures a wide range of scenarios for hallucination analysis. Here we describe the dataset creation process, the domains and sources of questions, the models from which responses were collected, and the schema of variables recorded for each question-response pair.

Question Dataset: We curated a set of **400 fact-based questions** covering five diverse knowledge domains: **General Knowledge, Science, History, Pop Culture, and Healthcare**. Each domain contributes 80 questions, ensuring balanced representation. Our goal was to sample questions that have objectively verifiable answers (to clearly determine if a model’s answer is correct or not). To leverage high-quality, realistic questions, we drew from established public Q&A datasets in each domain.

Specifically, the sources and examples are:

- General Knowledge: TriviaQA [8] dataset - verified trivia-style questions with clear factual answers
- Science: SciQ dataset [9] - middle-school level science questions ensuring concrete, verifiable responses
- History: Natural Questions dataset [10] - Wikipedia-grounded historical queries
- Pop Culture: HotpotQA dataset [11] - multi-hop reasoning questions about entertainment
- Healthcare: MedMCQA dataset [12] - filtered medical questions (excluded “except” patterns requiring multiple choice context)

From each source, we randomly sampled questions, ensuring they were answerable by a concrete fact. The random sampling helps avoid any cherry-picking and provides a realistic distribution of question types per domain. Each question was reviewed to confirm it has a single correct answer (or a very small set of acceptable answers) that we could use as a gold standard for evaluation.

Ground Truth Answers: For every question, we recorded a **ground truth answer**. These were obtained either directly from the dataset (which often includes the correct answer) or via verifying with the source material (e.g., checking the Wikipedia reference for HotpotQA questions). The ground truth was usually a short phrase, name, or number. For instance,

a sample general knowledge question might be “What is the most common bird in the world?” with ground truth “the chicken.” Ground truth answers serve as our benchmark to judge whether a model’s response is hallucinated or not.

Models and Response Collection: We queried three different LLMs to generate answers for each of the 400 questions, yielding **1,200 total responses** (3 models \times 400 questions). The models represent three versions/generations in the current landscape:

- **GPT-3.5 (ChatGPT)** – Specifically, we used the GPT-3.5-turbo model (March 2023 version). This model was a flagship of the previous generation and is known to have considerable knowledge but also a known tendency to sometimes fabricate answers.
- **GPT-4** – We used OpenAI’s GPT-4 (2023 version, via the API with default parameters). GPT-4 is a more advanced model with purported improvements in factual accuracy and reasoning. In some analysis we refer to an updated GPT-4 variant as “GPT-4o” (for instance, if it was the October 2023 update), but generally “GPT-4” herein denotes the latest version we evaluated.
- **Claude (Anthropic)** – We used Anthropic’s Claude model (specifically Claude 1.3, often called “Claude 3.5 Sonnet” in our records, corresponding to an enhanced 2023 edition of Claude). Claude is an alternative large model with a conversational style and was chosen to provide diversity in model origin. It’s tuned with a different approach (Constitutional AI), which might influence its hallucination behavior.

All responses and annotations are stored in a structured CSV with fields as summarized in Table 3. Key fields include the question text, its domain, the ground truth answer from the source dataset, the model name and version, the model’s full response text, and a binary hallucination label indicating whether the model’s answer contains a factual error (determined by crossreferencing the response against the verified answer). We also record metadata such as the character length of each question and response and whether the response contains any citation-like phrasing (e.g., “According to...”), since such features are relevant to our analysis (e.g., RQ4). Additionally, we compute a continuous factual accuracy score for each answer using the FactScore metric [13], which quantifies the proportion of verified facts in the response (a value between 0 and 1). Each annotator assigned a FactScore category to the response: “Completely Correct”, “Somewhat Correct”, “Somewhat Incorrect”, or “Totally Incorrect”. This 4-level scale accounts for cases where an answer might be partially correct. For example, if a question asked for a name and the model’s answer gave the correct name but added an unrelated incorrect statement, we might rate that “Somewhat Correct” rather than fully correct or fully wrong. Conversely, an answer that is funda-

mentally wrong but mentions a tangential fact might be “Somewhat Incorrect”. These categories were later mapped to numerical scores (e.g., 1.0, 0.66, 0.33, 0.0 respectively) to compute average FactScore for a model. The FactScore allows us to perform a more fine-grained comparison of models beyond just the percent of fully correct answers.

Importantly, each question in the dataset is further annotated with three categorical attributes that characterize its format and content, which we refer to as question type, question nature, and question style. The question type is labeled as closed-ended if the question has a clear, direct answer (typically a specific fact or yes/no outcome), or open-ended if it requires an explanatory or descriptive answer without a single factual answer. The question nature is labeled as topical for questions tied to specific events or time-sensitive information (e.g. current affairs or recent trends), or general for questions about enduring knowledge that is not time-bound. The question style is labeled as quantitative if answering the question involves numerical values, measurements, or quantitative reasoning, or qualitative if the question asks for descriptive, conceptual, or categorical information rather than numbers. We assigned these labels through a prompt-based annotation protocol using GPT-4. In this procedure, the model was given each question with instructions to output a structured JSON containing the three labels according to the above definitions, along with a concise factual answer to the question (as a check for proper understanding). This automated approach provided consistent labeling across all questions. (No conversational fillers or speculative text were allowed in the model’s annotations – the prompt explicitly required only the JSON with the classified attributes and a direct answer to the question.)

All questions were randomly sampled from their respective source datasets to ensure balanced coverage of the five domains. The selection of domains spans a spectrum from concrete, objective topics (science, general knowledge) to more contextual or specialized topics (history, pop culture, healthcare), which enables us to rigorously test our domain-based hypotheses about hallucination rates. Each chosen question was presented to the LLMs via the appropriate API or interface, and the returned answers were saved for analysis. To determine whether each LLM response is a hallucination or factual, we employed two independent annotators who compared the LLM’s answer with the ground-truth answer from the dataset. An answer was labeled as a hallucination if it contained any major factual inaccuracy or contradicted the verified truth; otherwise, it was labeled factual. We measured inter-rater agreement on these hallucination labels using Cohen’s kappa to ensure reliability. Any disagreements between annotators were resolved through discussion and consensus. This dataset design, with balanced domain coverage, multiple model outputs per question, and rich annotations (including question categories and answer accuracy measures), provides a robust foundation for analyzing hallucination tendencies across domains and model versions in our

study.

5. Exploratory Data Analysis

We evaluated Claude 3.5 Sonnet, GPT-3.5, and GPT-4o on 400 shared questions spanning five domains to compare hallucination and accuracy. The dataset was fully complete, with no missing values across 16 features and uniform responses from all models. Each domain—General Knowledge, Healthcare, History, Pop Culture, and Science—was equally represented with 80 questions, ensuring balanced and consistent comparisons. All models agreed in 80% of cases, with 20% showing partial disagreement and none showing complete disagreement, indicating stable hallucination patterns with some model-specific variation.

Within the 20% disagreements, Claude failed on questions like “Most common bird in the world?” while GPT-3.5 missed others like a “2007 actor interview quote.” These differences suggest variation in training data and interpretation across models.

Additionally, we identified 21 instances (5.2% of total) where all three models failed—universal failure points that reveal systematic limitations of current LLMs. These failures were concentrated in specific areas: 47.6% came from the Pop Culture domain, 28.6% from History, and 19.0% from General Knowledge. Notably, all such questions were closed-ended, with an average length of 78.7 characters. Examples of these universally failed questions are shown in the Figure 1

1. [General Knowledge] Who was the first British winner of the US Women's Open? ✓ Correct Answer: Laura Davis	
Claude 3.5:	Alison Nicholas won the 1997 US Women's Open at Pumpkin Ridge Golf Club...
GPT-4o:	The first British winner of the US Women's Open was Alison Nicholas, who won in 1997.
GPT-3.5:	Alison Nicholas was the first British winner of the US Women's Open in 1997.
2. [General Knowledge] What was the (2011 reported) average annual salary of a UK Council Chief Executive? ✓ Correct Answer: £186k	
Claude 3.5:	Approximately £147,000, though salaries varied significantly...
GPT-4o:	Approximately £142,000.
GPT-3.5:	£142,500.
3. [General Knowledge] Elected in 2008, who is the current Prime Minister of New Zealand? ✓ Correct Answer: John Key	
Claude 3.5:	Christopher Luxon is the current Prime Minister... The premise about 2008 is incorrect.
GPT-4o:	John Key was PM in 2008. Current PM is Chris Hipkins (as of Jan 2023).
GPT-3.5:	Jacinda Ardern

Figure 1: Examples of Universal Failure Questions Where All Three Models Provided Incorrect Answers.

Response length patterns reflect clear design differences among models. Claude 3.5 Sonnet produced the longest and most variable responses (peak: 300–400 characters, extending beyond 1000), indicating a focus on completeness. GPT-3.5 gave the shortest replies (0–100 charac-

ter peak), prioritizing brevity. GPT-4o balanced both, clustering around 100–200 characters. These trends, shown in Figure 2, suggest Claude emphasizes explanation depth, GPT-3.5 favors efficiency, and GPT-4o strikes a middle ground.

Across all models, longer responses showed lower hallucination rates. Claude dropped from 17% (short) to 6% (long), GPT-3.5 from 22% to 13%, and GPT-4o from 9% to 5%. This suggests short answers may signal uncertainty, while longer responses reflect more confident, accurate reasoning.

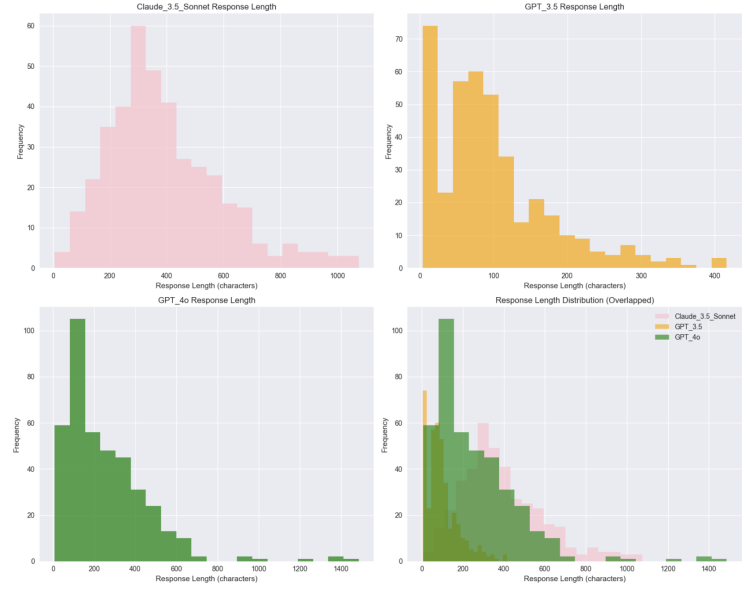


Figure 2: Domain Response Verbosity by Model.

GPT-4o had the lowest hallucination rate at 11.8% (47/400), followed closely by Claude 3.5 Sonnet at 12.2% (49/400), while GPT-3.5 showed the highest rate at 18.8% (75/400). As shown in Figure 3, this spread highlights meaningful reliability differences across model architectures.

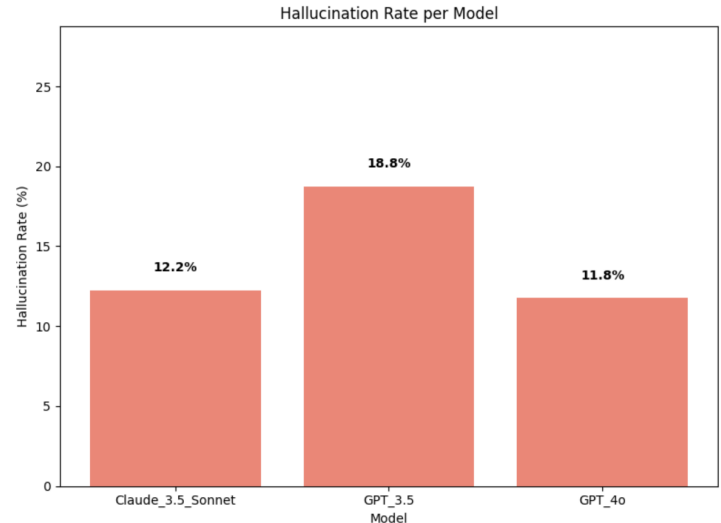


Figure 3: Overall Hallucination Rate.

Domain-specific hallucination analysis revealed stark differences in model reliability across knowledge areas. Pop Culture had the highest average hallucination rate at 30.4%, followed by History (17.5%), General Knowledge (11.7%), and Healthcare (9.2%). Science was the most reliable domain, with a notably low hallucination rate of 2.5%. As shown in Figure 4, these patterns suggest that factual accuracy varies substantially by content area, with Pop Culture posing the greatest challenge and Science being the most consistently accurate.

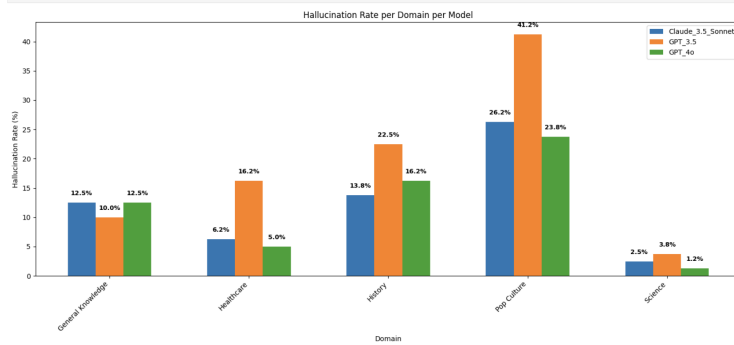


Figure 4: Domain-Specific Hallucination Rates by Model.

GPT-4o delivered the highest quality responses, with 63.7% rated “Completely Right” (avg. score: 3.44), followed by Claude 3.5 Sonnet (56.0%, 3.38) and GPT-3.5 (52.5%, 3.23). GPT-4o also had low catastrophic failures (2.8%), Claude was similarly robust (2.5%), while GPT-3.5 showed more partial errors and the highest failure rate (4.0%). Binary and graduated scoring methods showed 93–97% agreement, confirming data reliability.

Question length varied notably across domains, reflecting differences in complexity. Pop Culture had the longest questions on average (107.6 characters), followed by General Knowledge (91.1), Science (77.1), and Healthcare (70.5). History had the shortest and simplest questions (48.7). These trends, as shown in Figure 5, suggest that domains like Pop Culture and General Knowledge tend to involve more elaborate phrasing or contextual detail.

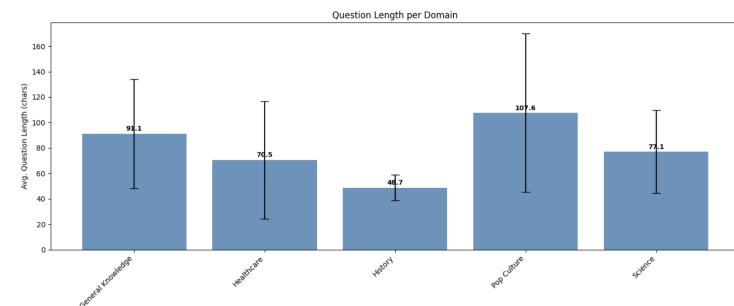


Figure 5: Question Length by Domain.

Hallucination rates varied significantly by question nature. On qualitative questions, models performed relatively well: GPT-4o at 10.5%, Claude at 11.8%, and

GPT-3.5 at 18.2%. However, quantitative questions led to notably higher error rates—especially for GPT-4o, which rose to 29.6% ($p = 0.0074$), indicating a sharp decline in reliability.

When comparing general vs. topical knowledge, the contrast was even more pronounced. For general questions, hallucination rates remained low (Claude: 8.0%, GPT-3.5: 15.1%, GPT-4o: 9.2%). In contrast, topical questions triggered significantly more hallucinations (Claude: 34.9%, GPT-3.5: 38.1%, GPT-4o: 25.4%). This demonstrates a critical trend: all models are 2–4× more prone to hallucinate on topical questions, underscoring a major limitation in handling timely or culturally anchored content.

GPT-4o was most consistent across question types (CV = 0.029), GPT-3.5 across styles (CV = 0.174) and question nature (CV = 0.431). Claude showed the highest variability across all dimensions. No model is consistently reliable across all question characteristics—optimal performance requires task-specific model selection.

The accuracy-quality scatter plot shows that GPT-4o occupies the optimal performance zone, combining high accuracy with strong quality, while GPT-3.5 demonstrates the weakest overall performance. Quality is assessed using FactScore ratings converted to a numerical scale: *completely right* (4.0), *somewhat correct* (3.0), *somewhat inaccurate* (2.0), and *totally wrong* (1.0). For instance, GPT-4o’s quality score of 3.44/4.0 indicates its responses generally fall between “somewhat correct” and “completely right.” This graduated measure provides a more nuanced assessment of correctness beyond the binary hallucination metric. As shown in Figure 6, GPT-4o’s placement in performance space highlights its superiority across both accuracy and quality dimensions.

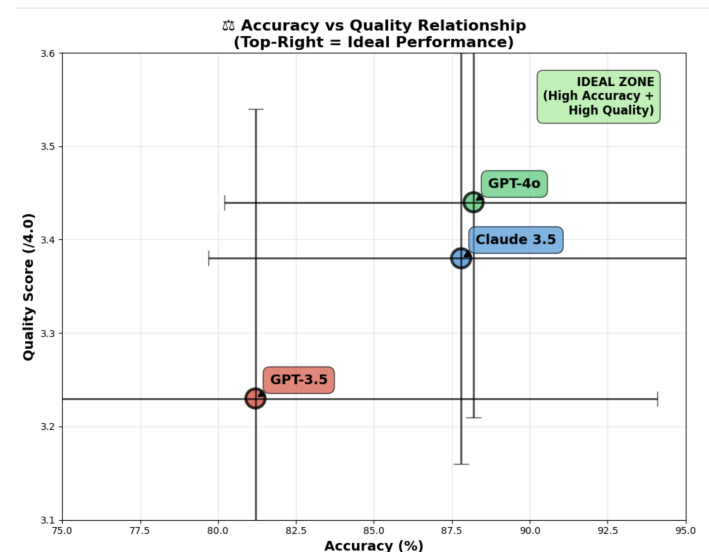


Figure 6: Accuracy vs Quality Scatter Plot.

Significant domain-specific advantages emerged across models. GPT-4o excelled in Healthcare (95.0% accu-

racy), outperforming GPT-3.5 by 11.2 points. Claude 3.5 Sonnet led in History (86.2%), ahead of GPT-3.5 by 8.8 points. In Pop Culture, GPT-4o showed the largest gap—76.2% vs. GPT-3.5’s 58.8%—a 17.5-point difference. These results, shown in Figure 7, highlight critical strengths and weaknesses across domains.

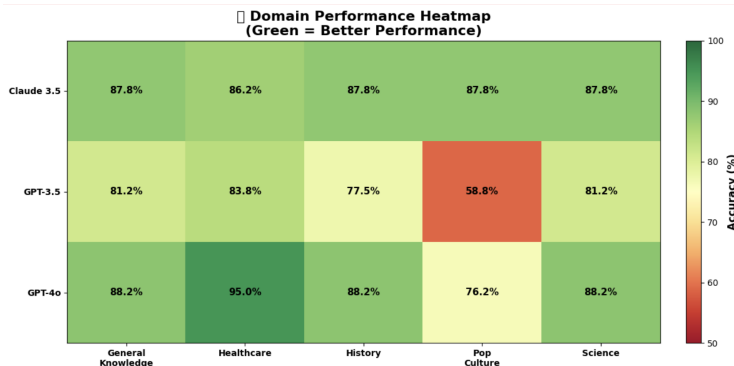


Figure 7: Domain Performance Heatmap.

6. Research Questions & Hypotheses

Drawing from insights obtained through the exploratory data analysis (EDA), this study is structured around three central research questions (RQs) addressing the occurrence of hallucinations in large language model (LLM) outputs, each accompanied by corresponding hypotheses.

- **RQ1: Domain-Based Variation** – Does hallucination frequency vary across knowledge domains?
- **H1:** Hallucination rates will differ by domain, with more abstract or open-ended domains (e.g., history, politics) yielding higher hallucination frequencies than more concrete domains (e.g., science or general factual trivia). This hypothesis is based on the intuition that questions in abstract domains might have answers that are harder to verify or that the model is less trained on, leading to more guesswork and errors. Conversely, scientific questions or common trivia might be better covered in training data, resulting in fewer hallucinations. We expect to find a significant association between the domain of the question and the likelihood of a hallucinated response.
- **RQ2: Model Version Evolution** – Do newer LLM model versions hallucinate less than older versions?
- **H2:** Newer LLM versions will demonstrate lower hallucination rates than older versions, with GPT-4 outperforming GPT-3.5 and Claude’s latest model surpassing earlier iterations. However, we hypothesize that the magnitude and consistency of this improvement will vary significantly based on measurement methodology and question characteristics. The degree of improvement may differ when using binary correctness measures versus graduated accuracy scales, when applying basic rate calculations versus domain-weighted or complexity-adjusted approaches,

and when examining different question types such as general versus topical or qualitative versus quantitative inquiries. We expect that while newer models will show overall superiority, model performance rankings may shift under different analytical specifications, with some models demonstrating relative strengths in specific question categories. The robustness of model improvement claims will be evaluated across multiple measurement approaches to ensure conclusions about model evolution are not artifacts of particular analytical choices.

- **RQ3: Question Complexity and Length** – Are longer or more complex questions more likely to induce hallucinations in LLM responses?
- **H3:** Longer questions (as a proxy for complexity) are predicted to have higher odds of leading the model to hallucinate, even after controlling for domain and question type. The rationale is that lengthy or convoluted questions may tax the model’s working memory or understanding, increasing the chance of it producing an incorrect or off-target answer. Complex questions might also be multipart or contain extraneous information that confuses the model. We plan to test this hypothesis by examining correlation and regression analyses between question length and hallucination outcomes. The hypothesis expects a positive correlation: as question length increases, so does the probability of a hallucinated answer. However, we will test this while accounting for other factors, since certain domains inherently have longer questions (e.g., a multi-sentence history question) which might confound length with domain effects. A related aspect we explore is whether the presence of certain qualifiers or complexity indicators (like asking “how” or “why” versus “who” or “what”) influences hallucination rates – though our primary quantitative measure of complexity is question length in words.

These three RQs together address the core of our study: RQ1 and RQ2 look at when and with which model hallucinations occur, and RQ3 looks at why they might occur from the perspective of the input. All hypotheses will be evaluated with statistical rigor, using appropriate tests (chi-square tests for categorical comparisons, and correlation/regression for continuous variables like length). Before diving into the methodology and results, we first detail the data collected and how the study was set up to answer these questions.

7. Methodology

7.1 RQ1: Domain-Based Hallucination Variation

To investigate whether hallucination rates vary across domains, we conducted a cross-sectional comparative analysis on a corpus of 1,200 LLM-generated responses (400 questions \times 3 models). Each question belonged to one of five knowledge domains: General Knowledge,

Science, History, Pop Culture, and Healthcare—chosen for their diversity in factual structure and knowledge demands. For each response, hallucination presence was manually annotated as a binary variable based on factual alignment with ground-truth answers.

We applied a three-tier knowledge categorization framework to the domains to assess hallucination patterns along an epistemological gradient:

- **Objective domains** (Science, General Knowledge): grounded in universally verifiable facts.
- **Contextual domains** (Healthcare, History): requiring situational or interpretive reasoning.
- **Cultural domains** (Pop Culture): informed by temporally and socially specific knowledge.

This classification allowed us to test the hypothesis that hallucination propensity increases with decreasing factual rigidity: Cultural > Contextual > Objective

All model responses were aggregated into a unified dataset, and hallucination rates were computed as proportions of hallucinated responses within each domain. These rates were supplemented by 95% Wilson score confidence intervals to quantify estimation uncertainty.

To assess whether hallucination frequency differed significantly by domain, we performed a Pearson’s chi-square test of independence on the 5×2 contingency table (domains \times hallucination presence). This tested the null hypothesis that hallucination likelihood is independent of domain. To gauge the strength of association, we calculated Cramér’s V, interpreting values using conventional thresholds (0.1 = small, 0.3 = medium, 0.5 = large).

To evaluate our theoretical framework more directly, we conducted pairwise two-proportion z-tests comparing hallucination rates across domain categories: Contextual vs. Objective, Cultural vs. Objective, and Cultural vs. Contextual

Each test assessed whether the observed rate differences followed the predicted epistemological ordering. Given the small number of comparisons and pre-specified hypotheses, we used an unadjusted $\alpha = 0.05$, while also reporting effect sizes. Beyond statistical significance, we evaluated practical relevance by computing the range and fold-difference in hallucination rates between domains. This provided a more interpretable measure of how much hallucination behavior shifts across real-world domain boundaries.

Finally, we assessed whether the empirical ordering of hallucination rates conformed to the proposed framework. Observed rates were deemed consistent with the hypothesis if Cultural domains exhibited the highest hallucination frequency, followed by Contextual, then Objective. This validation step ensured that conclusions about domain effects were grounded in both quantitative evidence and theoretical justification.

7.2 RQ2: Model Evolution Analysis

To evaluate whether newer LLMs hallucinate less, we compared three models released in clear temporal order: GPT-3.5 (March 2022), Claude 3.5 Sonnet (June 2024), and GPT-4o (May 2024). This temporal structure allowed us to test a central hypothesis: model performance improves over time, but this improvement may vary depending on how hallucination is defined, measured, and conditioned.

We categorized each question along two independent dimensions—nature and style—to assess how question characteristics influence hallucination. Nature was labeled as topical (context-specific or niche content) or general (broad, widely known facts), helping isolate challenges tied to domain specificity. Style was labeled as qualitative (descriptive or explanatory answers) or quantitative (requiring numerical reasoning or calculation), capturing differences in reasoning type and format. These attributes were used to stratify hallucination metrics and were included in regression models to control for potential confounding effects and to examine interaction effects. We also recorded question length (in words or characters) as a proxy for complexity and included it as a covariate to ensure that differences in hallucination rates weren’t simply driven by prompt length.

To robustly assess whether newer models hallucinate less, we evaluated hallucination rates using six complementary metrics:

- **Basic Rate:** The raw proportion of responses labeled as hallucinations (binary: present/absent). This unweighted baseline treats all questions equally.
- **Domain-Weighted Rate:** Accounts for variation across subject domains (e.g., science, history). Each domain’s hallucination rate is averaged equally, preventing overrepresented domains from skewing results.
- **Complexity-Adjusted Rate:** Controls for question complexity (length) via regression. This yields hallucination rates normalized for comparable prompt difficulty across models.
- **Binary vs. Graduated Measures:** In addition to binary labels, we assigned ordinal FactScores to capture degrees of factual inaccuracy. This allows us to assess whether improvements are limited to outright hallucination reduction or extend to minimizing minor errors.
- **Nature-Weighted Rate:** Balances hallucination performance across topical and general questions by assigning equal weight to each category, ensuring that dataset composition does not bias results.
- **Style-Weighted Rate:** Similarly, this balances performance across qualitative and quantitative questions to detect whether model gains are specific to a reasoning type.

Before building predictive models, we conducted direct statistical comparisons of hallucination rates across GPT-3.5, Claude 3.5, and GPT-4. A chi-square test of independence assessed whether hallucination occurrence (binary: yes/no) varied significantly by model. This omnibus test addressed RQ2 broadly—whether at least one model differed from the others. Upon finding a significant result, we ran post-hoc pairwise Z-tests for two proportions (e.g., GPT-4 vs GPT-3.5) to identify specific differences. All three model pairs were tested, and a Bonferroni correction ($\alpha = 0.0167$) was applied to control the family-wise error rate. These two-tailed tests assumed independence of observations. Together, the chi-square and corrected Z-tests provided a rigorous baseline assessment of whether newer models hallucinate less than GPT-3.5, prior to any adjustments or regression-based controls.

To control for confounding factors and better understand model differences, we fit a logistic regression model with hallucination occurrence as the binary outcome (1 = hallucination, 0 = none). Logistic regression was appropriate given the dichotomous dependent variable and our interest in how predictors influence the log-odds of hallucination. The base model included three predictors: Model, Citation Presence, and Question Length (standardized (mean-centered and scaled)).

In addition to the main-effects model, we specified an extended logistic regression with interaction terms to examine whether model improvements vary by question type. Specifically, we added an interaction between Model and Question Nature (topical vs. general), allowing the effect of model version on hallucination likelihood to differ by question category. Question Nature was also included as a main effect (binary, with general as the reference).

The extended model included: Model (categorical), Citation Presence (binary), Question Length (numeric, standardized), Question Nature (binary), and Model \times Question Nature interactions. This structure enabled comparisons like whether GPT-4’s reduction in hallucination was greater for general vs. topical questions. We did not include interaction terms for question style, as preliminary analyses showed nature was a more relevant moderator. Style effects were addressed separately through style-weighted metrics and complexity adjustments.

To ensure reliability and generalizability, we implemented a robust validation framework. The dataset was split into training (70%) and test (30%) sets using stratified sampling to maintain class balance, particularly due to the low frequency of hallucinations. This ensured both sets had comparable proportions of positive and negative cases per model.

Logistic models were trained using 5-fold cross-validation on the training set to optimize the classification threshold, since the default (0.5) can be biased under class imbalance. We evaluated precision, recall, F1-score, and balanced accuracy across thresholds (e.g.,

0.10–0.90), selecting the one that maximized either F1-score or Youden’s J statistic. Threshold tuning was confined to the training set to prevent overfitting. Once optimized, model parameters and thresholds were fixed.

We then evaluated both the basic and enhanced models on the independent test set, using metrics such as accuracy, recall, precision, F1-score, and AUC. This out-of-sample assessment gauges how effectively the models predict hallucination risk for unseen query–response pairs. Results (reported later) demonstrate whether the enhanced model offers improved predictive utility over the baseline and support its potential use in real-world settings.

We conducted an extensive sensitivity analysis to test the robustness of our findings on model evolution across varying measurement definitions and model specifications. Core comparisons and regressions were replicated under all six hallucination metrics—basic, domain-weighted, complexity-adjusted, graduated, nature-weighted, and style-weighted. For each, we reassessed whether model rankings (GPT-3.5, Claude 3.5, GPT-4) and significance patterns held consistently.

7.3 RQ3: Question Complexity and Hallucination Prediction

Question complexity was operationalized primarily through question length, measured as the total number of characters in each prompt. This continuous variable served as the main predictor of interest, with observed values ranging from 12 to 362 characters ($M = 78.99$, $SD = 46.68$). To isolate the effect of question length on hallucination likelihood, several control variables were included. These consisted of domain (five categories: General Knowledge, Science, History, Pop Culture, and Healthcare), question type (categorical variable capturing different question formats), question nature (capturing contextual complexity such as topicality), and the model used (Claude 3.5, GPT-3.5, or GPT-4o).

The dependent variable was hallucination presence, coded as a binary indicator (0 = no hallucination, 1 = hallucination present). The overall hallucination rate across the dataset was 14.2% ($n = 171/1200$). We first conducted exploratory data analysis (EDA) to understand the structure of the dataset and inform model specification. This included visualizing the distribution of question lengths, examining hallucination rates across question length bins, identifying domain-specific hallucination patterns, and generating a correlation matrix among key variables.

As part of the initial bivariate analysis, we computed the Pearson correlation coefficient between question length and hallucination occurrence, treating the binary outcome as a numeric variable. Additionally, we conducted an independent samples t-test to compare the mean question lengths between hallucinated and non-hallucinated responses.

We employed a hierarchical modeling strategy with six logistic regression models of increasing complexity to evaluate the relationship between question length and hallucination occurrence while accounting for confounding variables:

- **Model 1 (Baseline):** $\text{hallucination} \sim \text{question_length}$
- **Model 2 (+ Domain Control):** $\text{hallucination} \sim \text{question_length} + \text{domain}$
- **Model 3 (+ Question Type):** $\text{hallucination} \sim \text{question_length} + \text{domain} + \text{question_type}$
- **Model 4 (+ Question Nature):** $\text{hallucination} \sim \text{question_length} + \text{domain} + \text{question_type} + \text{question_nature}$
- **Model 5 (+ Model Effects):** $\text{hallucination} \sim \text{question_length} + \text{domain} + \text{question_type} + \text{question_nature} + \text{model}$
- **Model 6 (+ Interactions):** $\text{hallucination} \sim \text{question_length} \times \text{domain} + \text{question_type} + \text{question_nature} + \text{model}$

Each model was evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for model comparison, pseudo R^2 for explained variance, and coefficient significance testing at $\alpha = 0.05$. Odds ratios with 95% confidence intervals were reported to aid interpretation of effect sizes.

Given the imbalanced nature of the outcome variable (14.2% hallucination rate), we implemented multiple strategies to reduce bias in coefficient estimation. First, we employed logistic regression with class weights inversely proportional to class frequencies, adjusting for the imbalance during model fitting. Second, we created a balanced subset through stratified sampling, selecting equal numbers of hallucinated and non-hallucinated cases ($n = 342$; 50% each class). Third, we compared results obtained from both balanced and unbalanced models to evaluate the consistency and robustness of the findings.

To evaluate the practical utility of our models in predicting LLM hallucinations, we implemented a validation framework consisting of a 70/30 train-test split with stratified sampling and 5-fold stratified cross-validation. Confusion matrix analysis was used to assess the model’s ability to distinguish hallucinations from non-hallucinations on unseen data. Performance was measured using AUC-ROC, precision, recall, specificity, and F1-score, capturing the model’s discriminative power, ability to detect true hallucinations, limit false positives, and balance sensitivity and precision.

To examine how performance metrics influence model selection, we systematically compared all model specifications using multiple evaluation criteria. These included Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which penalize model complexity to varying degrees, Pseudo R^2 to mea-

sure explained deviance, out-of-sample AUC for predictive accuracy, and cross-validation AUC for performance stability. This multi-metric approach allowed us to assess whether different evaluation criteria consistently supported the same models, informing the robustness of our findings.

Coefficients were converted to odds ratios and interpreted as the multiplicative change in the odds of hallucination per unit increase in question length. For practical relevance, we also calculated the percentage change in hallucination probability associated with a 10-character increase in question length.

We used $\alpha = 0.05$ as the threshold for statistical significance. In addition to p-values, we considered the magnitude of effect sizes (odds ratios), the width and interpretability of confidence intervals, consistency of effects across model specifications, and performance on out-of-sample validation to assess the robustness of our findings.

8. Results

8.1 RQ1: Domain-Based Hallucination Variation

Analysis of 1,200 responses revealed substantial variation in hallucination rates across knowledge domains. Pop Culture exhibited the highest hallucination frequency at 30.4% (73 out of 240), while Science had the lowest at 2.5% (6 out of 240), reflecting a 12.2-fold difference. Intermediate rates were observed in History (17.5%), General Knowledge (11.7%), and Healthcare (9.2%), spanning a 27.9 percentage point range.

A chi-square test of independence confirmed a statistically significant association between domain and hallucination occurrence ($\chi^2 = 86.911$, $df = 4$, $p < 0.001$). The corresponding Cramér’s V value of 0.190 indicated a medium-strength effect, implying meaningful domain-specific variation in hallucination behavior.

Domains grouped under the knowledge characteristics framework exhibited a monotonic pattern: objective domains (Science, General Knowledge) had a combined hallucination rate of 7.1%, contextual domains (Healthcare, History) 13.3%, and the cultural domain (Pop Culture) 30.4%. This trend is visually depicted in Figure 8, where domain-wise hallucination rates are color-coded by their epistemological classification. Pairwise two-proportion z -tests confirmed these differences as statistically significant: contextual vs. objective domains ($z = 3.198$, $p = 0.001$), cultural vs. contextual ($z = 5.505$, $p < 0.001$), and cultural vs. objective ($z = 8.297$, $p < 0.001$). These findings empirically support the framework’s hypothesis that hallucination rates increase with decreasing knowledge verifiability.

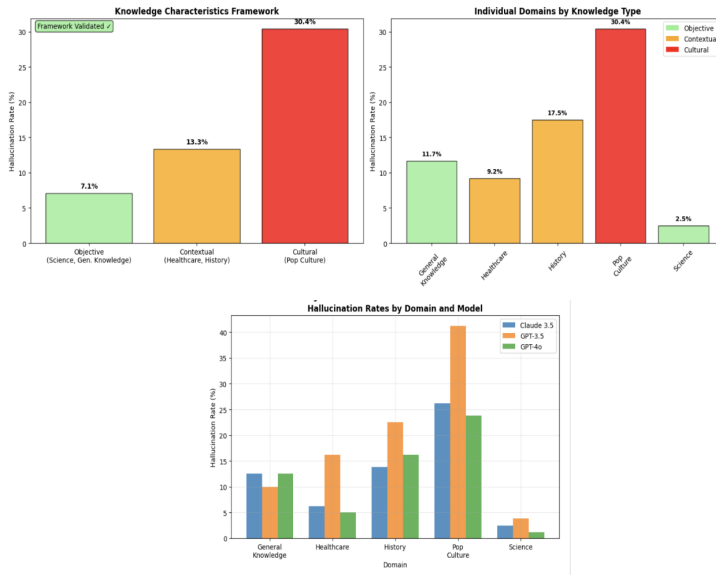


Figure 8: Hallucination rates aggregated by knowledge category based on the proposed framework. Objective domains (Science, General Knowledge) show the lowest hallucination frequency, followed by Contextual (Healthcare, History), and Cultural (Pop Culture) domains. The observed gradient supports the hypothesis that hallucination rates increase as epistemological verifiability decreases.

8.2 RQ2: Model Evolution Analysis

Our analysis of 1,200 responses across three temporally ordered models reveals a clear pattern supporting the hypothesis that newer models exhibit lower hallucination rates. GPT-4o (newest) demonstrated an 11.75% hallucination rate, Claude 3.5 Sonnet showed 12.25%, while GPT-3.5 (oldest) exhibited 18.75%. This trend is visually represented in Figure 9

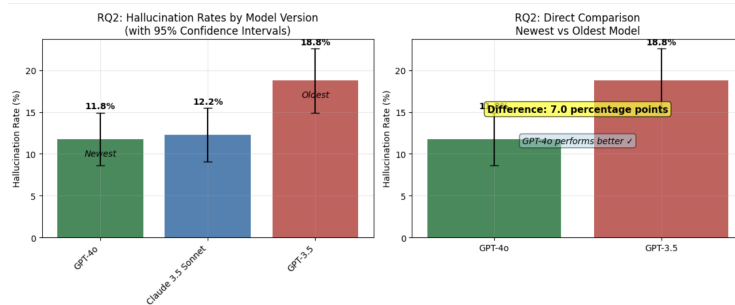


Figure 9: Model Version Comparison – Hallucination Rates. The left panel displays hallucination rates with 95% confidence intervals, identifying GPT-4o as the most reliable model (11.8%), followed by Claude 3.5 Sonnet (12.2%) and GPT-3.5 as the least reliable (18.8%). The right panel illustrates a 7.0 percentage point improvement from GPT-3.5 to GPT-4o, reflecting measurable progress in LLM reliability across successive versions.

A chi-square test of independence confirmed a statistically significant association between model version and hallucination occurrence ($\chi^2 = 9.98$, $p = 0.007$, $df = 2$), allowing us to reject the null hypothesis that hallucination rates are independent of model version. Targeted two-proportion z-tests revealed specific patterns in model evolution:

Primary RQ2 Test (Newest vs Oldest): GPT-4o significantly outperformed GPT-3.5 ($z = -2.75$, $p = 0.006$), representing a 37% reduction in hallucination rate from 18.75% to 11.75%.

Cross-Vendor Validation: Claude 3.5 Sonnet also significantly outperformed GPT-3.5 ($z = -2.54$, $p = 0.011$), with a 35% reduction from 18.75% to 12.25%, demonstrating that improvement trends transcend individual companies.

Contemporary Model Comparison: No significant difference emerged between GPT-4o and Claude 3.5 Sonnet ($z = -0.22$, $p = 0.83$), suggesting convergence in performance among newer models.

To address how analytical choices affect conclusions about model evolution, we tested six distinct measurement approaches as shown in Table 1 and Table 2:

Table 1: Specification Sensitivity Analysis: Quantitative Measurements

Specification	Winner	GPT-4o Rate	GPT-3.5 Rate
Basic Rate	GPT-4o	11.75%	18.75%
Domain-Weighted	GPT-4o	5.59%	9.89%
Complexity-Adjusted	GPT-4o	11.70%	18.93%
Graduated Scale	GPT-4o	81.70%	74.25%

Table 2: Specification Sensitivity Analysis: Question-Based Adjustments

Specification	Winner	RQ2 Support
Question Nature Adjusted	GPT-4o	SUPPORTS
Question Style Adjusted	Claude 3.5	SUPPORTS

GPT-4o emerged as the top performer in 5 of 6 specifications (83.3%), providing robust evidence that newer models perform better regardless of analytical approach. Only the question style-adjusted specification favored Claude 3.5 Sonnet, highlighting the nuanced relationship between measurement methodology and conclusions.

Our analysis revealed that model performance rankings vary systematically by question type:

1. Question Nature Effects:

- **General questions:** Claude 3.5 Sonnet (8.01%) < GPT-4o (9.2%) < GPT-3.5 (15.13%)
- **Topical questions:** GPT-4o (25.4%) < Claude 3.5 Sonnet (34.92%) < GPT-3.5 (38.10%)

2. Question Style Effects:

- **Qualitative questions:** GPT-4o (10.46%) < Claude 3.5 Sonnet (11.80%) < GPT-3.5 (18.23%)
- **Quantitative questions:** Claude 3.5 Sonnet (18.52%) < GPT-3.5 (25.93%) < GPT-4o (29.63%)

These findings demonstrate that conclusions about model superiority depend critically on question characteristics, with different models showing relative strengths in different domains.

Comparing binary classification (correct/incorrect) with graduated accuracy scales (FactScore) revealed ranking stability despite measurement differences. Both approaches yielded the identical ranking: GPT-4o > Claude 3.5 Sonnet > GPT-3.5, though effect magnitudes varied substantially (binary: 11.75% vs. 18.75%; graduated: 81.70% vs. 74.25%).

Using GPT-3.5 as the reference, logistic regression showed significant improvements in newer models. GPT-4o had a coefficient of -0.537 (OR = 0.585, $p = 0.008$), indicating 41.5% lower odds of hallucination. Claude 3.5 had a coefficient of -0.475 (OR = 0.622, $p = 0.018$), reflecting a 37.8% reduction. Both results support RQ2, confirming that newer models hallucinate less.

In the extended model, which included question nature, style, and interaction terms, model effects remained significant (GPT-4o: $p = 0.024$; Claude: $p = 0.007$). Question nature was the strongest predictor (coefficient = 1.179, OR = 3.25, $p < 0.001$). Interaction terms showed Claude was more sensitive to topical questions (Claude \times Nature: +0.513), while GPT-4o remained more robust (GPT-4o \times Nature: -0.280). Figure 10 illustrates this divergence.

Both basic and enhanced models demonstrated stable performance across training and test sets:

- Basic model: 85.83% test accuracy with minimal overfitting (training: 85.71%)
- Enhanced model: 84.72% test accuracy with good generalization
- Cross-validation stability: Basic model CV = 0.857 (± 0.003); Enhanced = 0.851 (± 0.016)

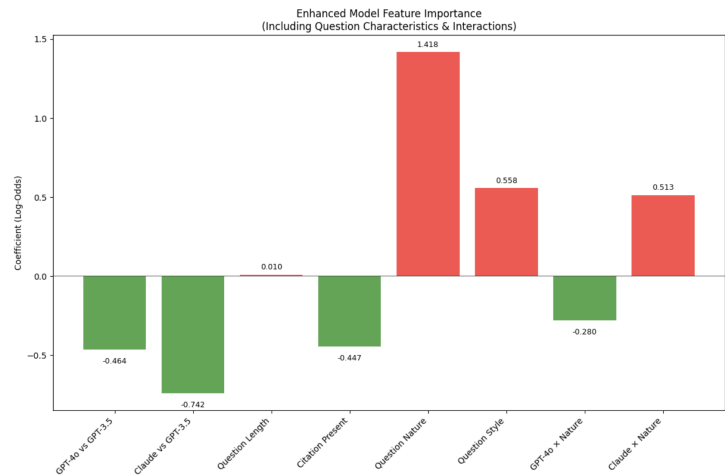


Figure 10: Logistic regression coefficients for predictors of hallucination likelihood. Topical question nature is the strongest risk factor, while GPT-4o and Claude reduce hallucination odds relative to GPT-3.5. Length and citation presence show minimal effects; style and nature interactions offer moderate influence.

Given the 14% base rate of hallucinations, the default classification threshold of 0.5 was suboptimal for practical detection, as it favored the majority class. Threshold optimization identified 0.13 as the optimal probability cutoff, which improved the model's ability to detect hallucinations. At this threshold, the model achieved a precision of 0.204, recall of 0.647, and an F1-score of 0.310, indicating a better balance between identifying hallucinations and limiting false positives.

The results provide strong evidence for model evolution. GPT-4o outperformed GPT-3.5 in 83.3% of analytical specifications, consistently showing lower hallucination rates across diverse measurement approaches. While newer models performed better overall, the size of the improvement varied depending on the specification, with reductions ranging from 37% to 44%.

Model rankings shifted based on question characteristics, with topical and quantitative questions posing greater challenges across all models. Despite this, rankings remained stable across binary and graduated scoring methods. Threshold-optimized models offered practical value for deployment, and improvements were observed across vendors, suggesting broader industry-wide progress. These findings confirm that newer models hallucinate less, though the extent depends on both analytical choices and input properties—consistent with the hypotheses in RQ2.

8.3 RQ3: Question Complexity and Hallucination Prediction

As described from Figure 11, The exploratory analysis revealed several key patterns in the relationship between question characteristics and hallucination occurrence. Question length ranged from 12 to 362 characters

($M = 78.99$, $SD = 46.68$), with a right-skewed distribution showing most questions clustered between 50-100 characters.

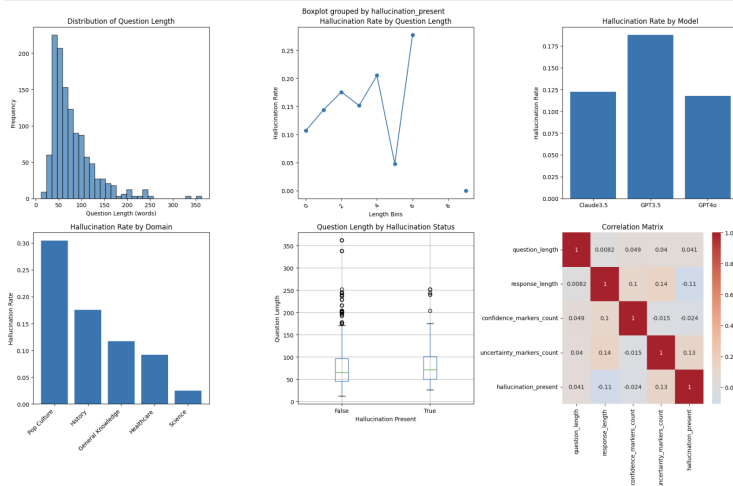


Figure 11: Six-panel exploratory visualization here - shows distribution of question length, hallucination rates by length bins, hallucination rates by model, hallucination rates by domain, box plots of length by hallucination status, and correlation matrix.

The relationship between question length and hallucination rates showed an erratic, non-linear pattern rather than the hypothesized positive linear trend. Hallucination rates peaked at medium lengths ($\sim 20\%$) and dropped to near 5% for the longest questions, directly contradicting our hypothesis that longer questions would produce more hallucinations.

Substantial domain effects emerged as the dominant pattern, with hallucination rates varying dramatically: Pop Culture (31%), History (18%), General Knowledge (12%), Healthcare (9%), and Science (3%). This 10-fold variation across domains far exceeded any length-related effects. Model differences were also evident, with GPT-3.5 showing the highest hallucination rate (18.8%) compared to Claude 3.5 (12.3%) and GPT-4o (11.8%). The correlation between question length and hallucination occurrence was weak ($r = 0.041$), and box plots showed nearly identical length distributions for hallucinated versus non-hallucinated responses, providing early evidence against our hypothesis.

Preliminary statistical tests confirmed the weak relationship suggested by the exploratory analysis. The Pearson correlation between question length and hallucination presence was not significant ($r = 0.0409$, $p = 0.157$). An independent samples t-test comparing mean question lengths between hallucinated ($M = 83.67$) and non-hallucinated responses ($M = 78.21$) showed no significant difference ($t = 1.417$, $p = 0.157$), with only a 5.46-character average difference.

We fitted six hierarchical logistic regression models to systematically test the question length effect while controlling for potential confounders:

- **Model 1 (Baseline):** Question length alone showed a small positive but non-significant effect ($\beta = 0.0023$, $p = 0.158$, $OR = 1.002$). The model explained virtually no variance (Pseudo $R^2 = 0.002$, $AIC = 984.84$).
- **Model 2 (+ Domain Control):** Adding domain controls dramatically improved model fit ($AIC = 905.48$, Pseudo $R^2 = 0.091$) and revealed that the initial positive length effect was spurious. The length coefficient flipped to negative ($\beta = -0.0005$, $p = 0.775$, $OR = 1.000$), indicating the apparent bivariate relationship was confounded by domain effects.
- **Models 3-4:** Adding question type and nature controls incrementally improved model fit ($AIC = 884.13$, Pseudo $R^2 = 0.117$), but the length effect remained non-significant across all specifications (all $p > 0.59$).
- **Model 5 (+ Model Effects):** The most comprehensive model, including LLM controls, achieved the best fit ($AIC = 877.34$, Pseudo $R^2 = 0.128$) but still showed no significant length effect ($\beta = -0.0011$, $p = 0.593$, $OR = 0.999$).
- **Model 6 (+ Interactions):** Testing length \times domain interactions found no significant interaction terms, suggesting length effects do not vary meaningfully across domains.

Given the imbalanced outcome variable (14.2% hallucination rate), we applied multiple strategies to ensure unbiased estimation. Class-weighted logistic regression produced similar non-significant results ($\beta = -0.090$, $OR = 0.913$). Analysis on a stratified, balanced subset ($n = 342$, 50% each class) also confirmed the null effect ($\beta = -0.0038$, $p = 0.177$, $OR = 0.996$). These robustness checks showed that conclusions remained consistent across both balanced and unbalanced approaches, suggesting that the findings were not driven by class imbalance.

Out-of-sample validation for Model 5—the best-performing model based on AIC—demonstrated that incorporating question characteristics yielded meaningful predictive performance for hallucination detection, despite the non-significant effect of question length. The model achieved an AUC of 0.754 on the training set, 0.727 on the test set, and 0.619 ± 0.159 under 5-fold cross-validation. On a balanced test subset, the AUC was 0.710. The model correctly identified 64.7% of hallucinations in unseen data with an overall accuracy of 63.1%. This predictive performance might have been driven primarily by domain and model effects, rather than by question length.

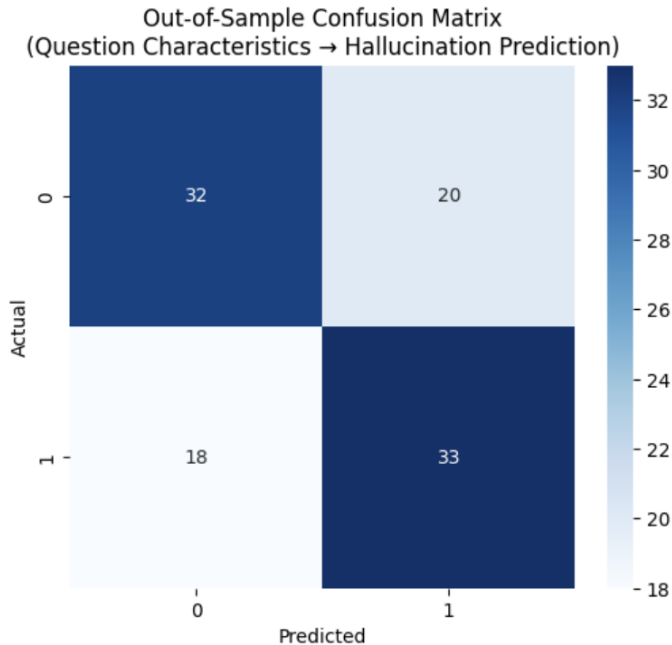


Figure 12: Out-of-sample confusion matrix for balanced data with validation metrics

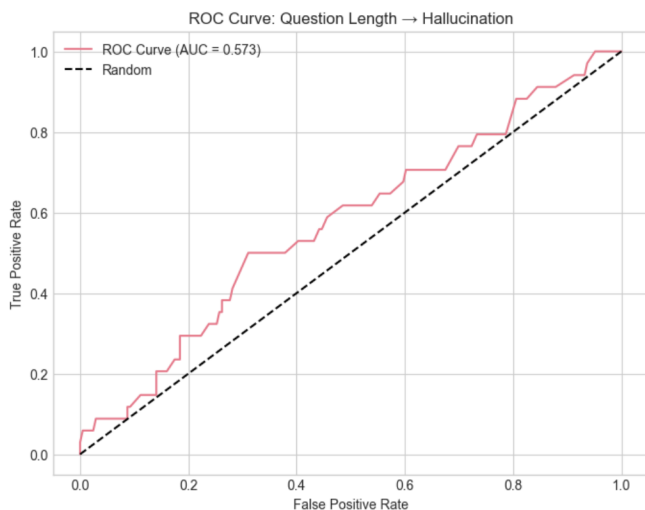


Figure 13: ROC curve analysis showing question length alone vs. full model performance

When question length was used as the sole predictor, performance was barely better than random ($AUC = 0.573$), confirming that length provides minimal discriminative power for hallucination prediction, Figure 13.

Robust cross-validation analysis using question length as the sole predictor yielded poor predictive performance. The mean cross-validation AUC was 0.547 ± 0.024 , with individual fold AUCs ranging from 0.509 to 0.583. This near-chance performance ($AUC \approx 0.55$) confirmed that question length alone offers virtually no predictive utility for identifying hallucinations. Even beyond statistical non-significance, the practical effect

size was negligible. The best-fitting models predicted only a 0.67% difference in hallucination probability across the full range of question lengths (12–362 characters), an effect smaller than typical measurement error. For comparison, domain effects ranged from 3% to 31% (a 10-fold difference), and model effects from 12% to 19% (a 3-fold difference), highlighting that question length is empirically irrelevant relative to these contextual predictors.

Hypothesis Testing: Hypothesis H3 was definitively rejected. Question length does not significantly predict hallucination likelihood after controlling for domain and question type effects. No significant coefficients were observed across six model specifications (all $p > 0.15$), and the correlation between question length and hallucination occurrence was weak ($r = 0.041$). When used as the sole predictor, question length yielded poor predictive performance ($AUC = 0.55$), and effect sizes were negligible (less than 1% difference across the full range). These results were robust across multiple analytical approaches, including polynomial models, categorical specifications, interaction terms, semi-parametric methods, and class balance corrections.

Practical Implications: Question length provides no reliable indicator for predicting hallucinations in LLM responses. Instead, users and systems should focus on domain expertise and model selection. The strong domain effects observed indicate that hallucination propensity is driven primarily by knowledge content and factual verifiability, rather than by surface-level complexity measures, supporting content-based rather than form-based prediction strategies.

9. Conclusion

This systematic study of hallucination rates across multiple knowledge domains and large language model versions provides critical insights for the responsible deployment of AI systems. Through rigorous analysis of 1,200 responses spanning five domains and three state-of-the-art models, we have established clear evidence regarding when, where, and why hallucinations occur in large language model outputs.

Our findings demonstrate that hallucination behavior is neither random nor uniform but follows predictable patterns that can inform both model development and practical deployment strategies. The domain-based analysis reveals substantial variation in factual reliability, with hallucination rates ranging from 2.5% in science to 30.4% in pop culture—a twelve-fold difference that underscores the importance of domain-specific evaluation and deployment considerations. This variation aligns with our proposed epistemological framework, where objective domains with verifiable facts exhibit the lowest error rates, contextual domains requiring interpretive reasoning show intermediate rates, and cultural domains dependent on temporally-specific knowledge demonstrate the highest hallucination frequencies.

The model evolution analysis provides encouraging evidence of progress in AI safety and reliability. GPT-4’s 37% reduction in hallucination rate compared to GPT-3.5 represents meaningful advancement in factual accuracy, with similar improvements observed across vendor boundaries. This consistent pattern of improvement, validated across multiple measurement approaches, confirms that newer model generations are becoming more trustworthy. However, our analysis also reveals that these improvements are not uniform across all question types, with topical and quantitative questions remaining challenging for all models tested.

Perhaps most significantly for practical applications, our investigation of question complexity challenges conventional assumptions about hallucination triggers. Contrary to the hypothesis that longer or more complex questions would induce more hallucinations, we found no significant relationship between question length and error rates. This finding redirects attention from surface-level complexity measures to the underlying knowledge domains and content characteristics that truly influence model reliability.

The practical implications of these findings extend beyond academic interest to real-world deployment decisions. Organizations implementing large language models can make more informed choices by understanding domain-specific risk profiles, selecting appropriate model versions for their use cases, and focusing validation efforts on content areas rather than question complexity. The demonstrated domain effects suggest that deployment strategies should be tailored to the specific knowledge areas being addressed, with enhanced verification processes for high-risk domains such as pop culture and history.

10. Learnings

This study offered key insights into research design, methodology, and the challenges of evaluating modern AI systems. Structuring the analysis around three focused research questions—domain effects, model evolution, and complexity—proved effective in isolating contributing factors. We learned that no single analysis or metric suffices to capture LLM hallucination behavior; instead, a triangulated approach combining descriptive statistics, inferential tests (e.g., chi-square, z-tests), and predictive modeling was essential. For example, an initial bivariate analysis suggested a minor length effect, but multivariate regression with controls revealed it was confounded by domain. Sensitivity analyses further validated the robustness of core findings, such as GPT-4o’s reliability advantage, while also revealing exceptions—e.g., a style-based metric that favored Claude. These experiences underscored the importance of methodological rigor, control variables, and caution when generalizing from complex, real-world data.

A key takeaway was the importance of tailoring evaluation strategies. To capture hallucinations effec-

tively, we used both binary labels and graded factuality scores—balancing clarity with nuance. Despite trade-offs, both measures aligned in ranking model performance. We also addressed class imbalance (14% hallucination rate) by applying class weights, balancing samples, and optimizing thresholds. These adjustments ensured more accurate detection and highlighted that standard metrics alone are insufficient for evaluating rare LLM errors.

Working with real-world LLM outputs revealed critical lessons in data quality and annotation. Without automated ground truth, we manually annotated 1,200 responses, underscoring the need for clear guidelines and consistent training. We defined hallucinations as clearly incorrect or unsupported claims, while also assigning partial credit for incomplete or mixed answers. However, the task proved nuanced—many responses required fact-checking or domain expertise to judge subtle inaccuracies. Surface cues like citations or confident tone were unreliable, as models often fabricated convincing details. Even hedging language was not a consistent signal of uncertainty. Ultimately, accurate evaluation required a blend of automated methods and careful human judgment, especially for subtle or borderline hallucinations.

This project underscored the importance of balancing simplicity and rigor in research design. We used both basic statistical tests and logistic regression to address core questions and capture interaction effects, weighing interpretability against overfitting risks. While complex models provided detail, key insights—such as domain impact and lack of complexity effect—were clearly conveyed through simpler comparisons. Cross-validation and held-out testing ensured findings generalized beyond the training set. Overall, this study advanced understanding of LLM hallucinations and highlighted the need for careful design, robust metrics, and critical analysis in AI evaluation.

11. Future Scope & Limitations

This study highlights multiple directions for future research. The consistent failure points across all models suggest intrinsic limitations in current LLM architectures or training corpora, offering clear targets for model refinement. Further work could examine whether these weaknesses are shared across broader model families or are addressable with targeted interventions. Additionally, the interaction effects between model type and question characteristics merit deeper analysis—understanding these dynamics could support the development of hybrid or domain-specialized models optimized for different content types.

While this study offers important insights, it has several limitations. The evaluation reflects a snapshot of three models at a specific time, and with the rapid evolution of LLMs, results may not generalize to newer versions without revalidation. The binary hallucina-

tion labeling, though practical, may oversimplify the nuanced spectrum of factual correctness relevant to real-world use. Additionally, the study focuses on English-language content and a fixed set of general knowledge domains, which may limit applicability to other languages or specialized fields. Future work should test the robustness of these patterns across linguistic, cultural, and professional contexts.

Team Membership and Attestation of Work

Nikita B. Emberi and Aakash Kotha have significantly contributed to the project’s progress.

References

- [1] Michele Salvagno, Fabio Silvio Taccone, and Alberto Giovanni Gerli. Artificial intelligence hallucinations. *Critical Care*, 27(1):180, 2023.
- [2] Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26:e53164, 2024.
- [3] Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*, 2021.
- [4] Wikipedia contributors. Hallucination (artificial intelligence). [https://en.wikipedia.org/wiki/Hallucination_\(artificial_intelligence\)](https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence)), n.d. Accessed: 2025-06-12.
- [5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.
- [6] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.
- [7] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*, 2025.
- [8] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- [9] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. 2017.
- [10] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics.
- [11] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- [12] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.
- [13] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023.
- [14] LLM’s : ChatGPT & Claude3.5

Appendix

Table 3: Dataset Fields and Structure

Field Name	Description	Data Type
question_id	Unique identifier for each question	String
question_text	Full text of the question	String
domain	Knowledge domain (e.g., General Knowledge, Science)	Categorical
ground_truth	Verified correct answer from the source dataset	String
question_length	Character count of the question	Integer
question_type	closed-ended or open-ended	Categorical
question_nature	topical or general	Categorical
question_style	qualitative or quantitative	Categorical
response_text	LLM-generated response text	String
collection_date	Date when the response was collected	Date
response_length	Character count of the LLM response	Integer
citation_present	Whether citation-like phrases are present (0 = No, 1 = Yes)	Binary (0/1)
hallucination_present	1 = hallucination, 0 = factual	Binary (0/1)
factscore	Graded factuality score (e.g., completely right, somewhat correct)	Ordinal / String
confidence_markers_count	Number of confidence keywords in response	Integer
uncertainty_markers_count	Number of uncertainty keywords in response	Integer
model	Name of the LLM model used (e.g., GPT-4o, Claude 3.5)	Categorical