
# LibreEval

## Phoenix Open Source Hallucination Evaluation Model & Dataset

Version 1.1

# Table of Contents

# Abstract

## Introduction

Retrieval Augmented Generation (RAG) is used in modern AI applications to enhance language models by integrating external knowledge retrieval. However, a persistent issue in applications that rely on this retrieval process is hallucination, where models generate responses that appear credible but are not grounded on the retrieved information. These hallucinations present significant challenges for real-world deployments, particularly in high-stakes domains such as healthcare, legal, and finance, where misinformation can have serious consequences.

Despite growing research efforts to understand and mitigate LLM hallucinations, existing studies often focus on isolated aspects such as hallucination detection, causal analysis, or mitigation strategies. Some of the most performative models for detection of hallucinations are proprietary and incur higher costs than open-source models to deploy in production applications.

Our motivation for this work stems from the need for more transparent, systematic, and scalable approaches to hallucination detection and mitigation in RAG applications. Specifically, we sought to:

1. **Provide stronger evaluators for identifying hallucinations in RAG applications:** While LLMs have shown remarkable capabilities, their hallucination tendencies make them unreliable for many critical tasks. We aim to offer more precise evaluation tools that can reliably detect and categorize hallucinations.

2. **Create a flexible and extensible platform for preparing RAG hallucination datasets:** Researchers and practitioners often require domain-specific datasets to train models tailored to their needs or evaluate their own applications. We aim to allow users to generate their own RAG hallucination datasets from specified web domains, ensuring adaptability across various applications.

3. **Support tuning and evaluation using both custom and prebuilt datasets and models:** Users can either bring their own data to evaluate models and leverage LiteLLM to test against a wide array of models.

To address these gaps, we introduce **LibreEval**, an open-source platform that enables systematic generation, evaluation, and benchmarking of RAG hallucination datasets. Alongside the platform, we release LibreEval1.0, **the largest open-source RAG hallucination dataset to date**. The LibreEval1.0 dataset consists of 74,917 samples consisting of 13,899,271 tokens, and is made up English (52,946), Portuguese (9,739), Japanese (8,165), Korean (1,432), French (1,255), Spanish (1,240), and Chinese (140), and samples spanning the domains of Finance, Technology, Health, Business, Science, and Law. We also release fine-tuned GPT-4o-mini and Qwen2-1.5B-Instruct models that were tuned on the LibreEval1.0 dataset and that improve detection of hallucinations compared to their base model counterparts.

# Related Works

Recent research has made significant strides in understanding and mitigating hallucination in large language models (LLMs). Multiple studies have explored both the underlying causes of hallucination and the development of robust evaluation methodologies, while also emphasizing the critical role of diverse training data. For instance, HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models (https://arxiv.org/pdf/2305.11747) introduced a comprehensive dataset of 35,000 generated and human-annotated samples, revealing that nearly 20% of model responses contain hallucinations. Its successor, The Dawn after the Dark: An Empirical Study on Factuality Hallucination in Large Language Models (https://arxiv.org/pdf/2401.03205)(HaluEval 2.0), extends this work by categorizing hallucinations into fine-grained subtypes—such as "outdated information hallucinations" versus "unverifiable information hallucinations"—and by rigorously examining underlying causes like insufficient or imbalanced training data. These findings underscore the need for dataset diversity across multiple domains, including finance, law and medicine.

Parallel efforts have focused on developing efficient methods for hallucination evaluation. For example, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (https://arxiv.org/pdf/2306.05685) demonstrated that an LLM judge can agree with human preference decisions over 80% of the time—comparable to inter-human agreement levels. This has motivated the development of fine-tuned, smaller LLM judge models that improve on cost and latency. PandaLM (https://arxiv.org/pdf/2306.05087) introduced a fine-tuned LLaMA-7B judge model capable of evaluating instruction-following abilities through pairwise comparisons. Later, JudgeLM (https://arxiv.org/pdf/2310.17631) achieved state-of-the-art performance by matching GPT-4's evaluations over 90% of the time, surpassing typical human-human agreement levels. More recently, Halu-J (https://arxiv.org/pdf/2407.12943)was released as a 7B judge model specifically designed to detect hallucinations, offering detailed critiques rather than simple factuality scores. Together, these models present a cost-effective alternative to relying on large-scale models like GPT-4 or on human annotators.

Despite their strong performance on in-domain tasks, small judge models face limitations in broader generalization. Studies such as An Empirical Study of LLM-as-a-Judge for LLM Evaluation (https://arxiv.org/pdf/2403.02839) highlight that achieving robust generalization necessitates training on large and diverse datasets. Synthetic data generation techniques have thus garnered considerable attention as a means to produce such datasets. Approaches like Self-Instruct (https://arxiv.org/pdf/2212.10560): Aligning Language Models with Self-Generated Instructions and Alpaca (https://crfm.stanford.edu/2023/03/13/alpaca.html): A Strong, Replicable Instruction-Following Model have shown that effective instruction tuning can be achieved using fully synthetic datasets. These methods substantially reduce the need for costly human annotation while enhancing data diversity and coverage, ultimately improving model robustness by exposing them to a broader range of scenarios.

Collectively, these works reveal several key insights: (1) hallucinations often stem from limitations in training data diversity and quality; (2) robust evaluation frameworks—supported by benchmarks such as HaluEval, HaluEval 2.0, MT-Bench, and Chatbot Arena—are essential for reliably measuring hallucination, though current benchmarks still have room for improvement in capturing nuanced error types and domain-specific challenges; (3) synthetic data strategies, as exemplified by Self-Instruct and Alpaca, offer scalable, cost-effective alternatives to human annotation; and (4) while small fine-tuned judge models provide significant benefits in cost and latency, their effectiveness hinges on training with large, diverse datasets to ensure sufficient generalization. These converging lines of research provide the backdrop for this paper, which aims to develop more reliable and versatile LLM benchmarks alongside open-source, fine-tuned LLM evaluators.

# LibreEval1.0 Dataset

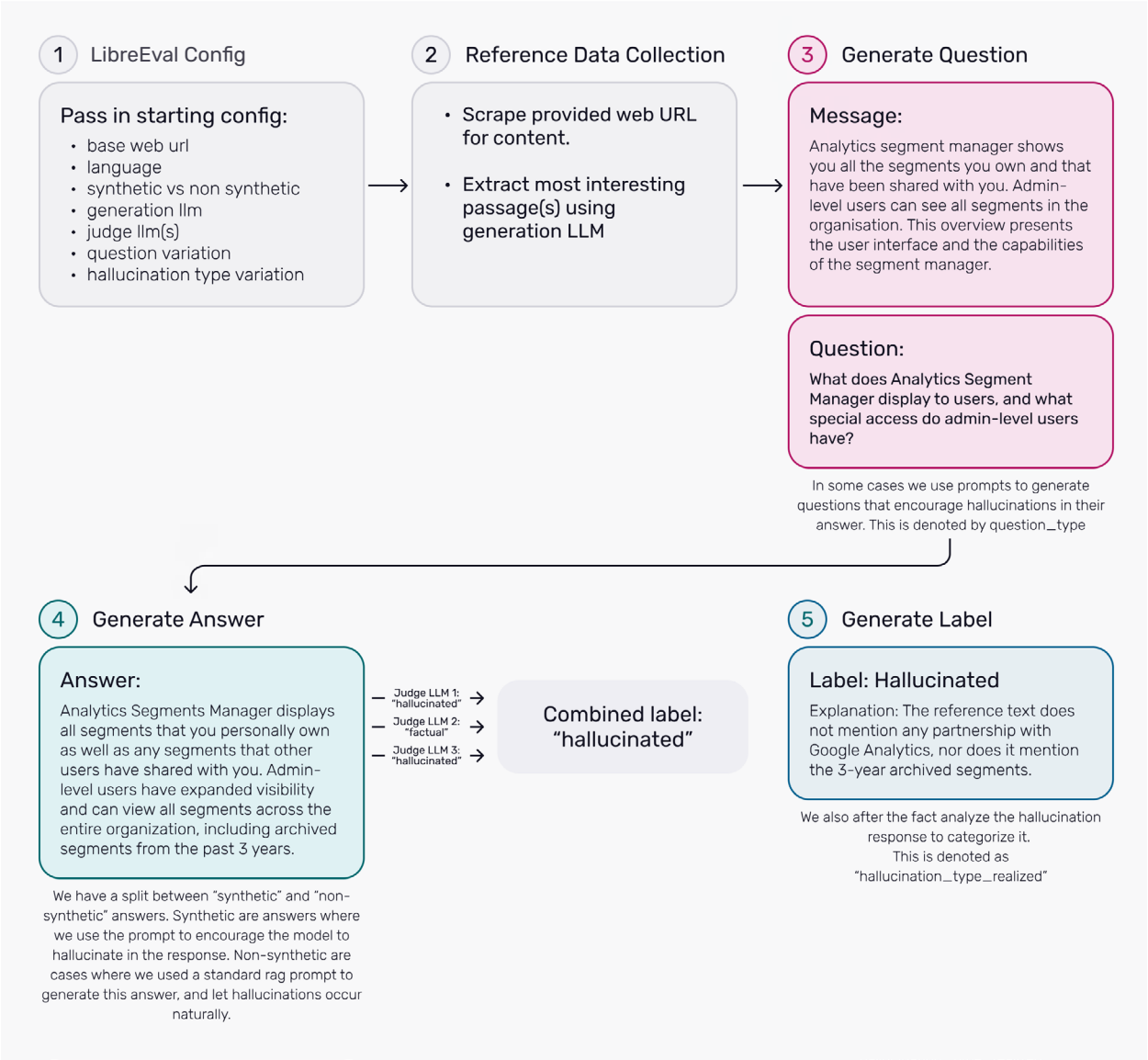
We release LibreEval, an open source framework for generation, evaluation, and benchmarking of RAG hallucination datasets. To promote transparency and reproducibility, the repository used in this research has been fully open-sourced, providing a foundation for further expansion. Given token and latency constraints, we focus exclusively on one-shot tuning data, optimizing for efficiency without sacrificing quality.

## Dataset Format

Our methodology constructs a dataset by pairing user input questions with relevant reference document context, sourced from web scraping, and corresponding LLM-generated responses. This approach ensures the dataset closely reflects a typical RAG application, capturing diverse question types and response structures to effectively train hallucination evaluators.

| Term | Definition |
|---|---|
| Input/Question | The question posed by a user to a RAG application |
| Context/Reference | Reference materials relevant to a user's question |
| Output/Response | The LLM generated answer based upon the input, context, and prompt |
| Label/Evaluation | The boolean label "Hallucinated" or "Factual" |
| Explanation | A paragraph describing the justification for the boolean evaluation outcome |

## Dataset Generation Workflow




The LibreEval framework enables rapid dataset creation for evaluating and finetuning hallucination models. Here are the steps:

1. Configs Setup - The configurations of the dataset generation depend on the dataset source, language, what kinds of questions are more relevant to the user, hallucination types, synthetic or non-synthetic datasets. These can all be defined and configured based on the user's dataset needs.

2. Reference Data Collection: The reference data used in RAG will be scraped and the most interesting or relevant passages will be extracted. These references should be similar to the user's intended dataset to maximize performance of the finetuned model.
3. Question Generation: There are different types of questions that can be generated based on user configuration. In some cases, we use prompts to generate questions that encourage hallucinations in their answers.
4. Response Generation: We have a split between "synthetic" and "non-synthetic" answers. Synthetic answers are where we use the prompt to encourage the model to hallucinate in the response. Non-synthetic are cases where we used a standard rag prompt to generate this answer, and let hallucinations occur naturally.
5. Label Generation: The hallucination labels are generated using a council of judges, and the majority is the final label.

## Reference Data Collection

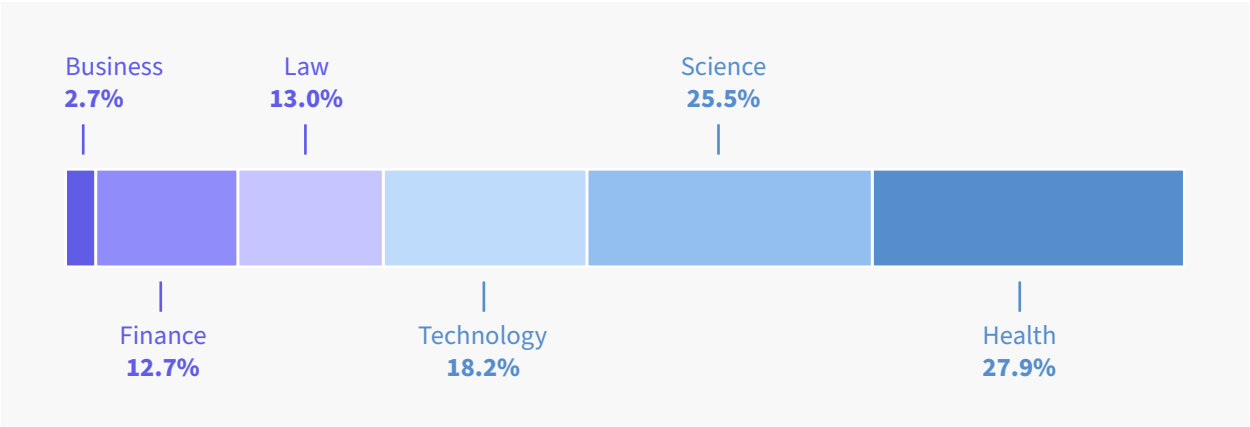
To generate LibreEval1.0, a diverse collection of datasets was compiled from various sources spanning web scrapings between November 2024 to January 2025. Domains were chosen to give a diverse set of facts and technical instructions to which questions could be generated. The diversity was important to potentially capture data that had not yet been used to train existing LLM models.

Domains for the English dataset were selected across a range of topics including widely used technical resources (technology 18.2%) alongside critical fields like health (27.9%) and science (25.5%), which require factual precision. The inclusion of finance (12.7%), law (13.0%), and business (2.7%) provides specialized knowledge, helping to train LLM hallucination evaluators on complex, high-stakes topics where accuracy is essential.

We also generated a multilingual dataset for finetuning models. We have included the distributions in the "Multilingual Dataset Distributions" section of the appendix for researchers interested in using this data. Due to the significant class imbalance of the dataset favoring English data sources, our finetuned models are focused on the English dataset.

| Website | Domain | Description |
|---|---|---|
| Investopedia | Finance | A financial education website offering articles, investment guides, and market analysis. |
| Databricks Documentation | Technology | A technical resource for cloud-based big data analytics and machine learning. |
| MongoDB Documentation | Technology | Official documentation for MongoDB's NoSQL database and related technologies. |
| NOAA Research | Science | A government agency providing research and data on climate, oceans, and atmospheric science. |
| NASA Earth Observatory | Science | A NASA website featuring satellite imagery and articles on Earth's climate and environment. |
| Cornell Law School (LII) | Law | A legal resource offering access to U.S. laws, case law, and legal interpretations. |
| NCBI | Health | A database of biomedical research, genetics studies, and scholarly articles. |
| PMC (PubMed Central) | Health | A repository of open-access biomedical and life sciences journal literature. |
| MedlinePlus | Health | A public health resource providing consumer-friendly medical and drug information. |
| Adobe Experience League | Business | A business and marketing resource for Adobe software documentation and best practices. |

**Distribution of Knowledge Domain in English Data**



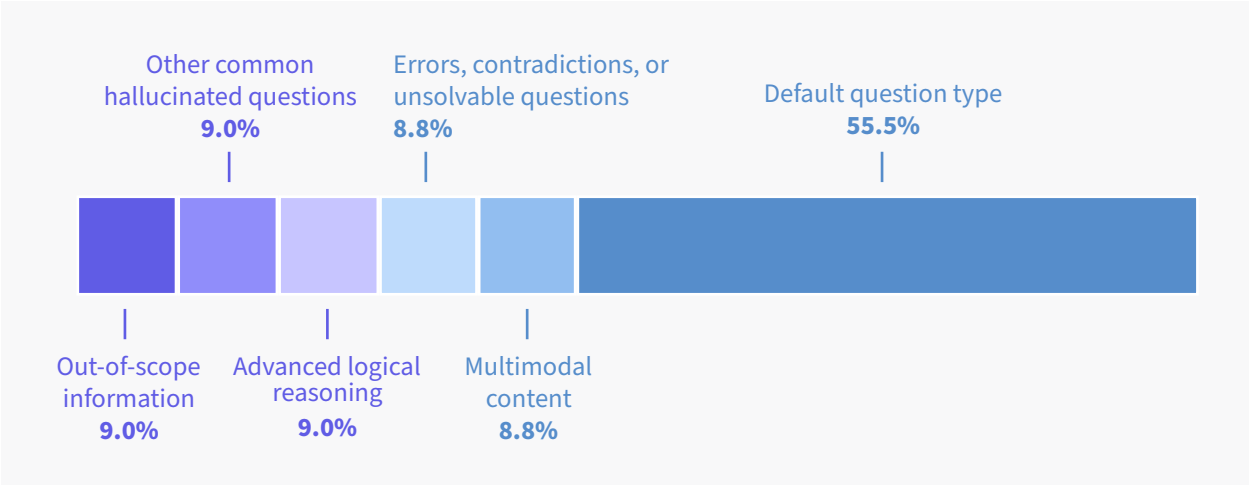

## Question Generation

We simulated the typical workflow of a RAG application, which combines user queries with relevant context. First, we divided the reference data into semantically similar chunks. We identified a range of question types that would pose distinct challenges to an LLM, and created prompts based upon these categories of questions to best simulate the diverse range of questions asked of LLMs in real world data. In some cases we did not prompt our model to generate questions of a specific category, for which we classified these as “Default Question Type”. Questions were generated using the models: *GPT-4o, Claude-3.5-Sonnet, and Llama-3.1-8b.*

Our efforts aimed to balance evaluation by combining realistic user queries with challenging edge cases that test an LLM’s reasoning and factual accuracy. The default question type (55.6%) provides a baseline, while categories like advanced logical reasoning (9.5%) and errors, contradictions, or unsolvable questions (8.4%) push the model’s limits. By incorporating out-of-scope (9.1%), multimodal (8.9%), and hallucination-prone (8.5%) questions, this dataset effectively simulates real-world interactions while identifying potential failure points in model responses.

## Distribution of Question Types

### English Only - Distribution of question_type



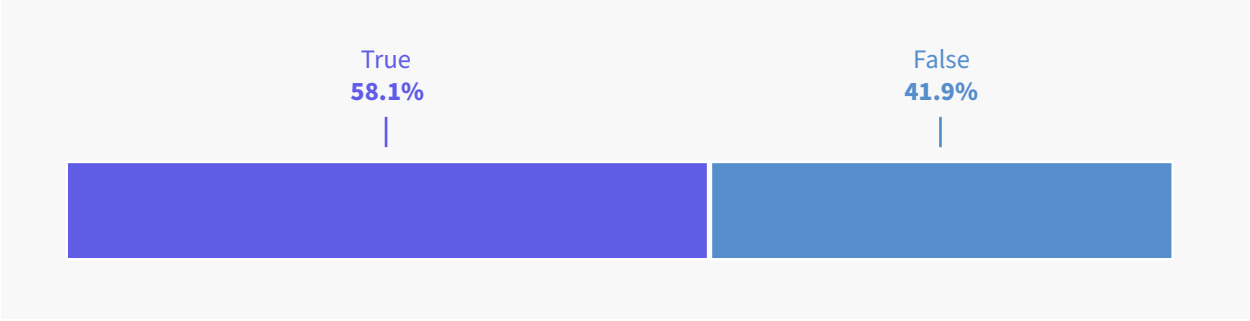

## Response Generation

Our efforts focused on creating a robust dataset of hallucinated data to finetune a model capable of detecting various types of hallucinations in RAG applications Naturally occurring hallucinations may be limited to only a few categories of hallucinations, making it difficult to train a well-rounded model on this dataset. To address this, we incorporated synthetic hallucination data to ensure coverage across a wide range of hallucination types. By strategically generating different types of hallucinations, we aimed to prevent class imbalance, ultimately enabling more robust hallucination detection. Our approach leveraged a majority of synthetic data (59.0%) to ensure comprehensive coverage of hallucination types that might not naturally occur in sufficient quantities, preventing dataset imbalance and enhancing hallucination detection models.

To generate responses, we provided question and context pairings to an LLM, which was prompted to answer the input question. The LLMs used for this task included GPT-4o, Claude-3.5-Sonnet, and Llama-3.1-8b. In some cases, the prompt instructed the LLM to make a sincere effort to answer using the provided RAG context, producing what we classify as non-synthetic data. In other cases, the data generation script randomly selected a prompt directing the LLM to hallucinate, generating responses classified as synthetic data. This approach allowed us to diversify the dataset while avoiding biases that could arise from relying solely on naturally occurring hallucinations. Below, we detail the nature of these prompts and our methods of response generation.

| Definitions | Description |
|---|---|
| **Synthetic (Hallucination) Data** | Data generated by prompting the LLM to hallucinate intentionally. These responses are created using specific prompts designed to induce different types of hallucinations, allowing us to cover a diverse range of hallucination types and balance the dataset. |
| **Non-Synthetic (Hallucination) Data** | Data generated by prompting the LLM to answer questions sincerely using the provided RAG context. These responses may still contain hallucinations but are not intentionally induced; rather, they emerge naturally when the model attempts to answer with limited or misleading context. |

**English Only - Distribution of Synthetic**




## Label Generation

The generated responses were compared with input and context pairings to identify hallucinations, which were defined as instances when a response contained information not represented in the input question or context. This was performed using a LLM-as-a-judge strategy, where each response was analyzed by a council of LLMs using a standard prompt (https://github.com/Arize-ai/phoenix/blob/21ca64be24c3b069478e714d0d6a1936e54aa761/packages/phoenix-evals/src/phoenix/evals/default_templates.py#L82) to identify hallucinations and each judge marked a response as "factual" or "hallucinated". The council consisted of GPT-4o, Claude-3.5-Sonnet, and Qwen-2.5-7b and the “ground truth” label assigned to each hallucination evaluation was selected by majority vote. Additionally, all judges generated explanations justifying their label.

| Example Data Inserted into a Prompt and a Generated Response | |
|---|---|
| **Context / Reference:** | MongoDB is the most widely available, globally distributed database, capable of running anywhere. MongoDB Atlas offers a fully managed experience, available in over 115 regions across major cloud providers. It enables elastic scaling, allowing you to adjust capacity up or down as needed, all without having to manage infrastructure. |
| **Input / Question:** | Data generated by prompting the LLM to answer questions sincerely using the provided RAG context. These responses may still contain hallucinations but are not intentionally induced; rather, they emerge naturally when the model attempts to answer with limited or misleading context. |
| **Output / Response:** | Based on MongoDB's aggressive expansion strategy, they plan to add approximately 40 new regions by 2025, primarily focusing on emerging markets in Southeast Asia and Africa. The company has also announced partnerships with Oracle Cloud and IBM Cloud to expand beyond their current cloud provider partnerships, though these rollouts have been slightly delayed due to regulatory challenges in certain markets. They're particularly targeting data sovereignty requirements in India and Brazil, where they aim to establish 8-10 new regions within the next 18 months. |
| **Generated Label:** | Hallucinated |

The data was also labeled for the type of hallucination generated by a response. A council of LLM judges was used for this, consisting of GPT-4o, Claude-3.5-Sonnet, and Qwen-2.5-7b.

## Human Labels

Preceding data deduplication, the dataset was sent to the human labelling provider LabelBox for manual annotation of hallucinated data.

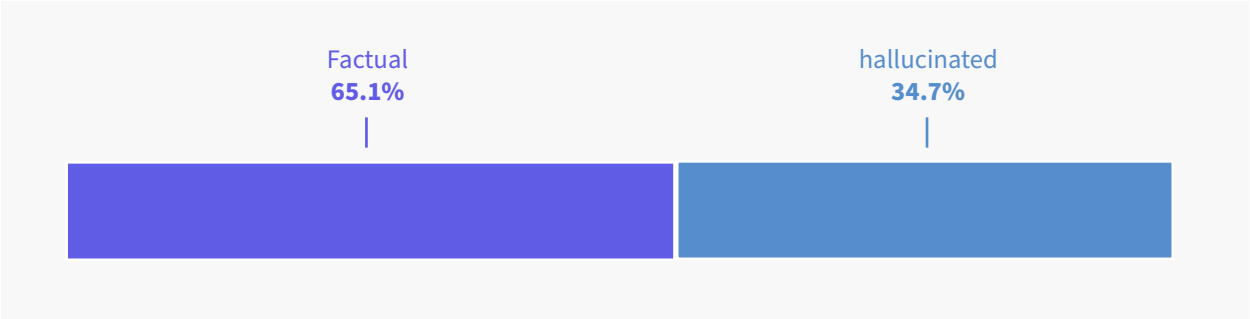
Two instruction sets were tested in order to verify the accuracy of the human labels. First, the labelers were presented with both the task and ensemble of judges majority label. In this first iteration, the human annotators were tasked with verifying the correctness of the judge labels.

The second test removed the ensemble of judges response from the data given to the human labelers.

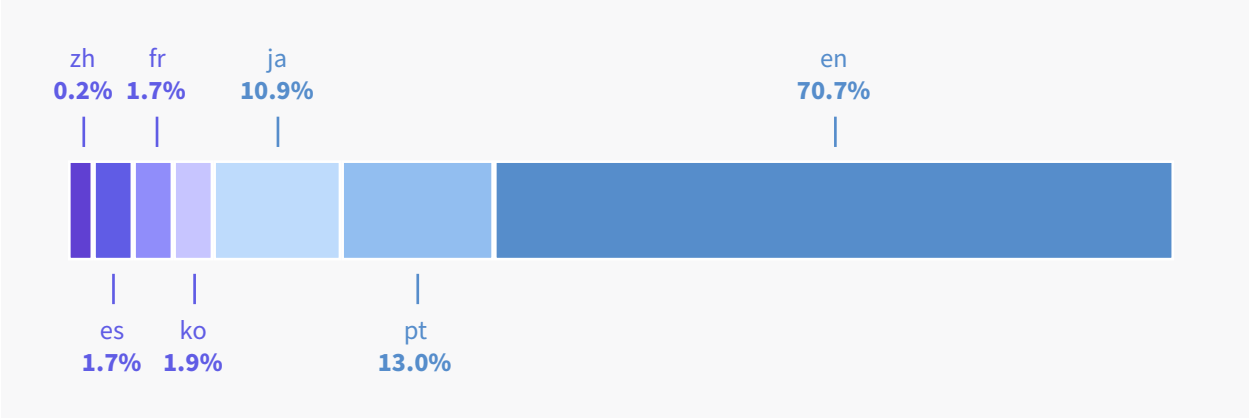
## Resulting LibreEval1.0 Dataset

The resulting dataset consists of 74,917 labeled examples. All rows contain three LLM judge labels and explanations, and 15,024 of these rows also contain human labels. Due to the stronger performance of the council of LLM judges compared to the human labelers, the ultimate “label” column is decided by the council majority vote, and does not factor in the human labels (https://arxiv.org/pdf/2306.05685).

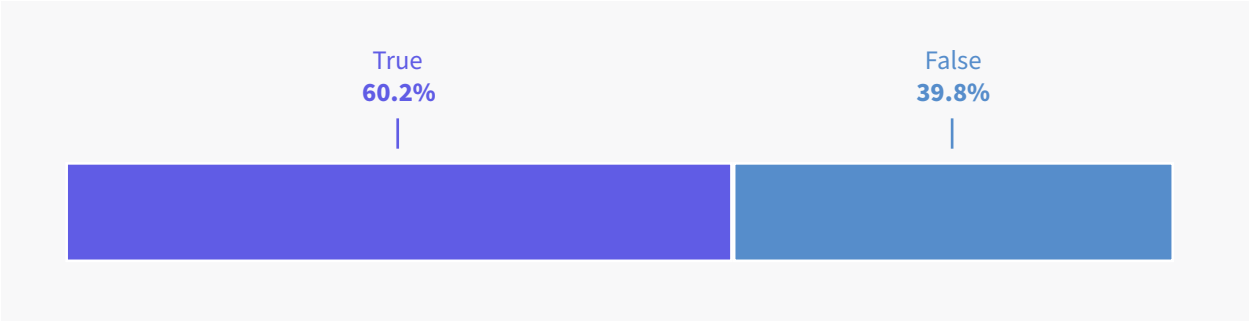
### All Combined - Distributon of label



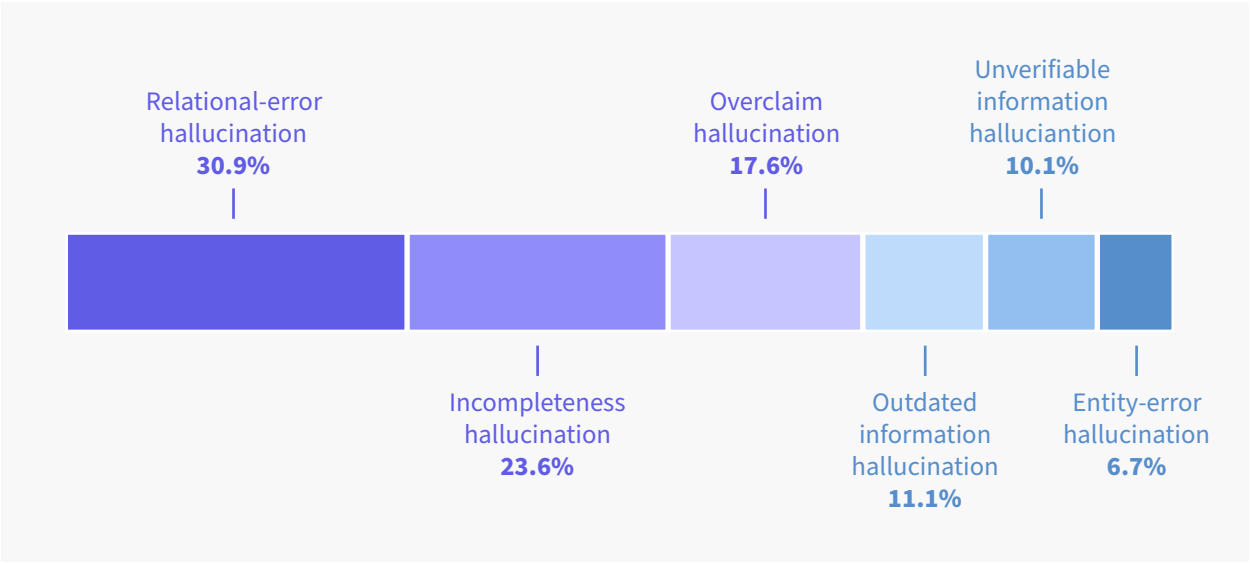

### All Combined - Distributon of language



**All Combined - Distributon of synthetic**




**All Combined - Distributon of hallucination_type_realized**




For more details on the insights from the generated datasets and labeling process, see the “Datasets” section in the Appendix.

# LibreEval1.0 Fine Tuned Hallucination Models

## Data Preparation for Fine-Tuning Hallucination Evaluation Models

We release two Phoenix Eval hallucination detection models, trained by fine-tuning GPT-4o-mini and Qwen2-1.5B-Instruct models. The data used for training consisted of examples from data sources with English language content. Data was deduplicated on exact matches between input, reference, and output rows. We applied a train (70%) / validation (15%) / test split (15%) to our datasets.

**English Dataset Synthetic and Hallucination Sample Counts for All Data**

| Synthetic | Factual | Hallucinated | Total |
|---|---|---|---|
| Synthetic | 15,079 | 15,600 | 30,679 |
| Non-Synthetic | 20,181 | 1,984 | 22,165 |
| Total | 35,260 | 17,584 | 52,844 |

## Models Fine-Tuned

We fine-tuned two variants of Qwen/Qwen2-1.5B-Instruct models using the Together.ai platform. The training datasets were provided via training and validation JSONL files in messages format. We used LoRA (Low-Rank Adaptation), and our configuration employed a rank (lora_r) of 8, a scaling factor (lora_alpha) of 16, and was applied to all linear layers (lora_trainable_modules = "all-linear"). To prevent gradient explosion, we applied gradient clipping (max_grad_norm = 1). The training was conducted on 8 GPUs per node, with automatic batch size allocation. The training process was initiated from epoch 0 with no prior weight modifications. Full hyperparameters for these training runs can be found in the appendix, and a summarized view of selected different parameters can be found below.

We fine-tuned the GPT-4o-Mini-2024-07-18 model using supervised fine-tuning on a dataset structured in a prompt-response format. The training and validation data were provided in JSONL one-shot message format. Fine-tuning was performed for one epoch, processing a total of 39,049,341 tokens. A batch size of 8 was used, and the learning rate was scaled by a factor of 0.3 relative to the base model's default learning rate.

We offer 3 fine-tuned models, two variants of Qwen2 1.5b Instruct and a GPT 4o Mini fine-tune. The primary differences between the model variations are as follows:

| Model | Prompt structure | Hyperparams |
|---|---|---|
| GPT 4o mini Fine-tune | Eval prompt supplied as a system prompt. Row data supplied as a user prompt | Batch Size: 8<br>Learning rate multiplier: 0.3<br>Epochs: 3 |
| Qwen v1 Fine-tune | Eval prompt supplied as a system prompt. Row data supplied as a user prompt | Batch size: 8<br>Learning rate: 3e-06<br>Warmup ratio: 0.4<br>Epochs: 3 |
| Qwen v2 Fine-tune | Full prompt, including row data, supplied as a system prompt | Batch size: 32<br>Learning rate: 3e-05<br>Warmup ratio: 0.1<br>Epochs: 3 |

## Evaluation of Models

Base and fine-tuned models were deployed as endpoints on Together AI compute (hardware: RTX6000-48GB) and OpenAI compute respectively. Evaluations were performed using the LibreEval repository published with this research. Evaluations were run both using the test split of LibreEval as well as outside datasets HaluEval 1.0 and the Fever, HotpotQA, NQ, and WoW datasets used in ARES. HaluEval 1.0 was prepared by mimicking the approach used in the original repository - typically by pulling a correct or hallucinated answer as the listed output for a row at equal rates. ARES datasets were used as is, with the Answer_Faithfulness_Label mapping to the ground truth label.

| Dataset | Example Count |
|---|---|
| HaluEval 1.0 - QA | 10,000 |
| HaluEval 1.0 - Dialogue | 10,000 |
| HaluEval 1.0 - Summarization | 10,000 |
| HaluEval 1.0 - General | 4,507 |
| HaluEval 1.0 Total | 34,507 |
| ARES - HotpotQA | 11,200 |
| ARES - Fever | 20,888 |
| ARES - NQ | 5,306 |
| ARES - WoW | 6,108 |
| ARES Total | 43,502 |
| LibreEval | 3,444 |
| **Total Overall Eval Examples** | **81,453** |

Evaluation of model performance was performed with the LibreEval1.0 library. The evaluation methods consisted of inserting input, context, and output data into the "HALLUCINATION PROMPT TEMPLATE" prompt template available through the arize-phoenix repository on github (https://github.com/Arize-ai/phoenix/blob/21ca64be24c3b069478e714d0d6a1936e54aa761/packages/phoenix-evals/src/phoenix/evals/default_templates.py#L720). The LLM judge council consensus label was used as a ground truth for evaluation purposes. LibreEval was passed "--evaluation-models" and "--dataset-to-evaluate" arguments to initiate an evaluation run that called the llm_classify function available through the Phoenix open source package. Responses were then snapped to either a "factual" and "hallucinated" label.

# Results

Initially we set out to achieve three goals:

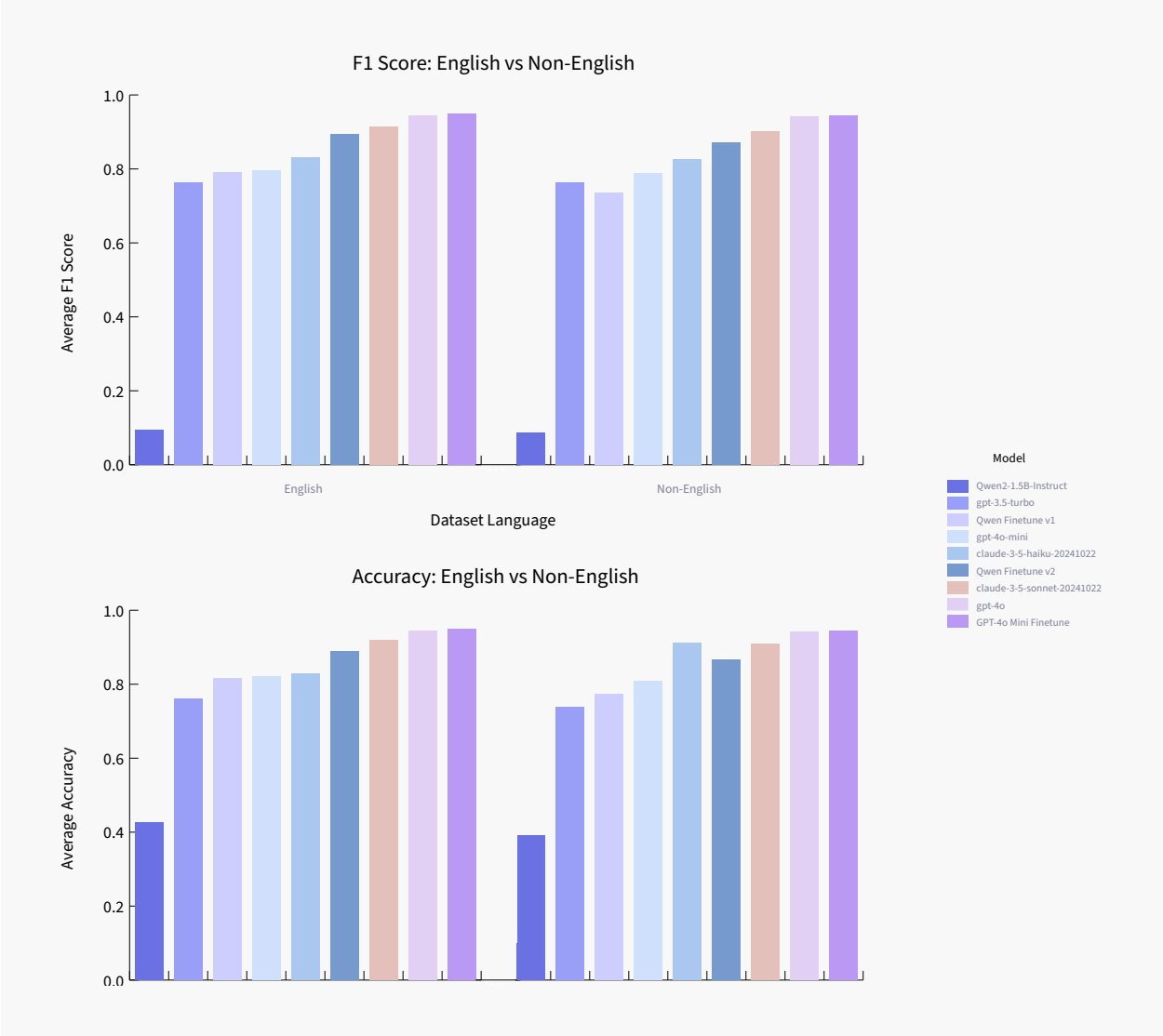
1. Create an open source repo that can be used to generate new labeled hallucination evaluation datasets, using supplied text as a corpus.
2. Generate a large, diverse dataset of hallucinations across different domains, languages, and types.
3. Train a set of <1.5b parameter models that perform at comparable levels to existing SLMs, and a fraction of the cost.

We believe we've achieved each of these goals.

All of the code used for this project is available in the LibreEval repository. The code has been designed to not only be reproducible, but extensible to not only generate additional datasets, but to evaluate existing models on different datasets as well.

All of the datasets generated are in the same LibreEval repository and available on Hugging Face.

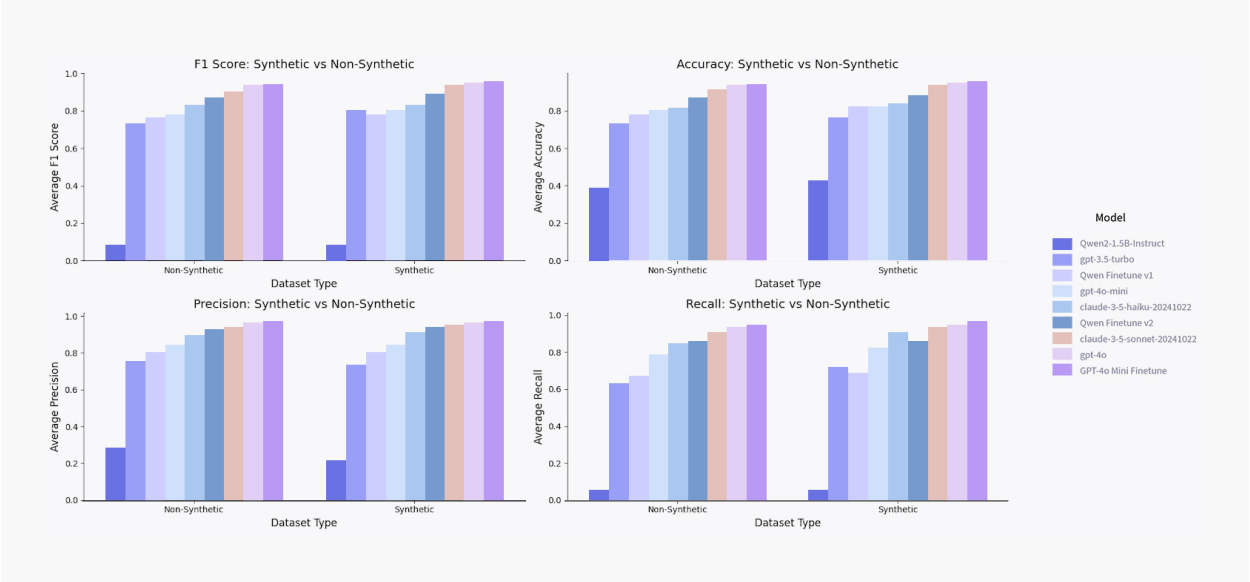
## Model Evaluation Performance Against LibreEval1.0 by Language




English vs other languages did not appear to substantially change performance for any particular model. The largest difference came in the first Qwen fine-tune: 0.056 better performance in f1 score when tested solely on english examples.

## Model performance of Base and FT Models by Synthetic vs Non-synthetic

Another split within the generated datasets is whether some type of encouragement to hallucinate was used in the generation process. In some cases, the model was encouraged to hallucinate through language in its prompt. These cases are referred to as “synthetic” hallucinations.

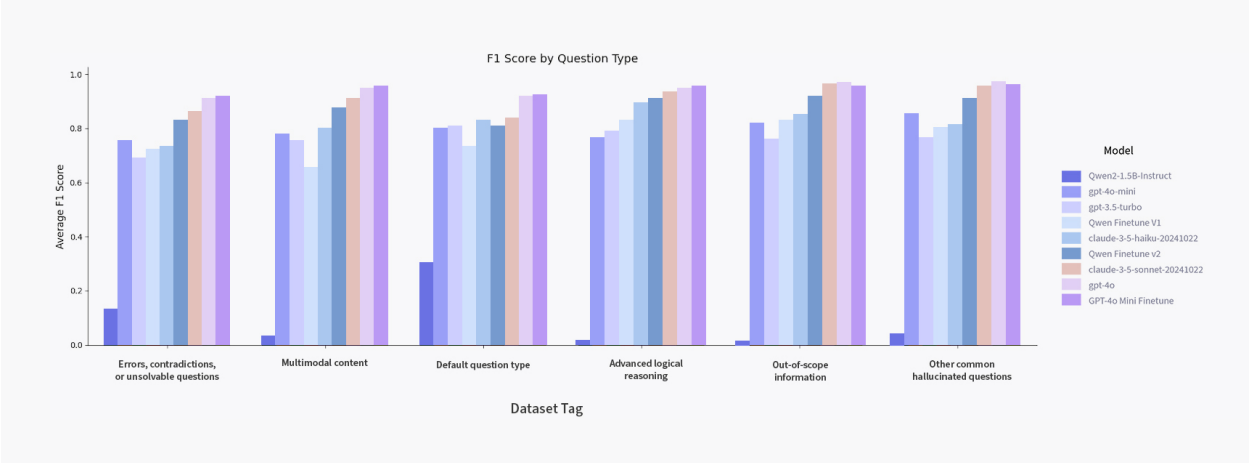


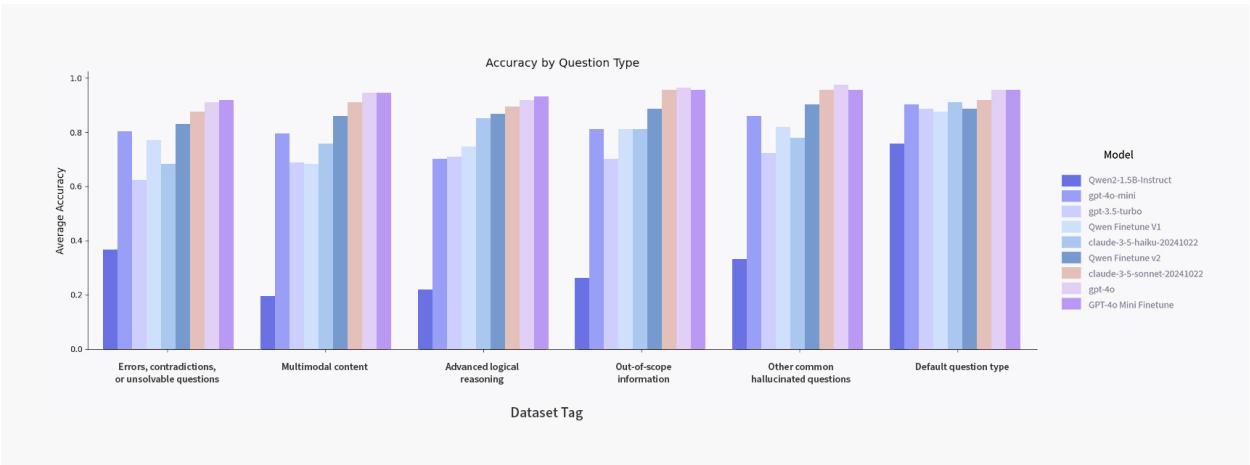

Looking at model performance across these synthetic hallucinations vs non-synthetic hallucinations, it’s notable that existing base models have an average lower recall of 22.20% on non-synthetic cases. In other words, base models miss a larger number of non-synthetic hallucinations than they do synthetic hallucinations.

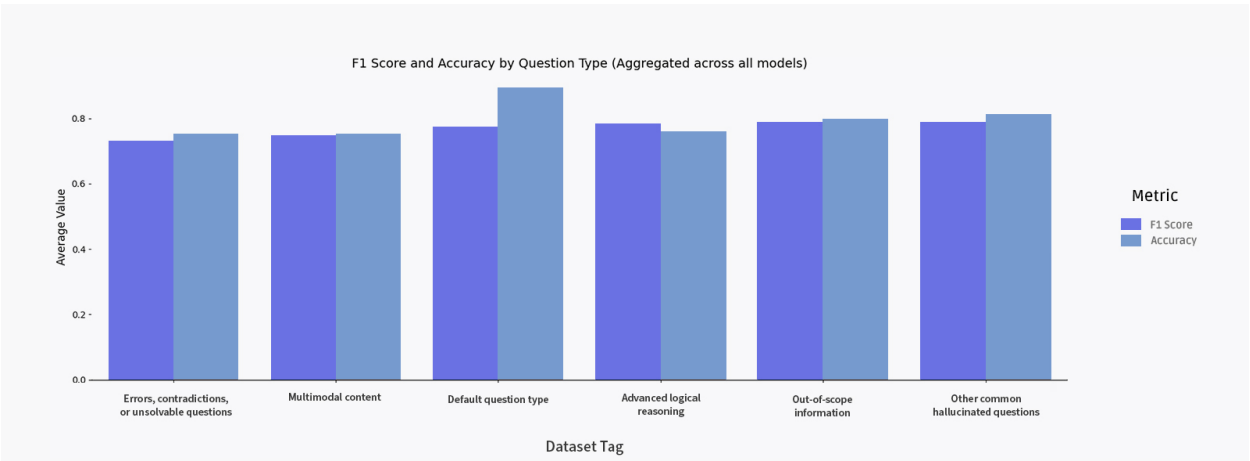
This is notable, as some existing hallucination evaluation datasets rely on synthetically generated hallucinations, which could overrepresent a model’s performance.

## Model performance of Base and FT Models by Question Type

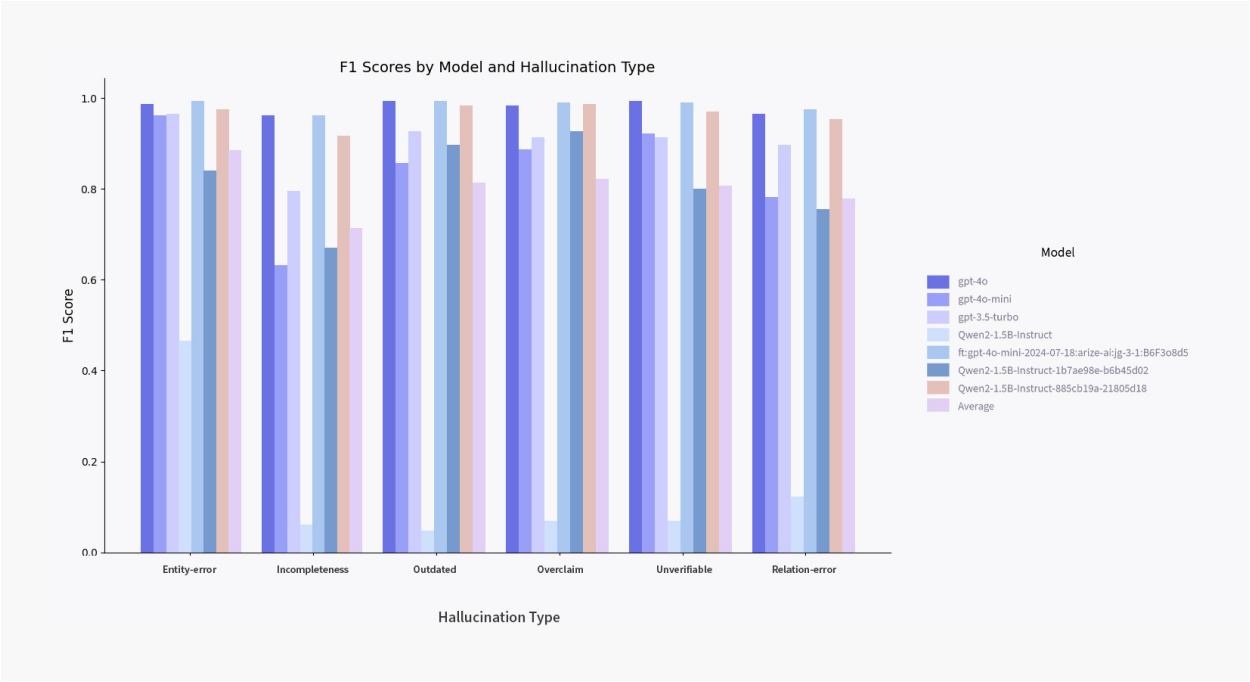
Another variation of the generated datasets was on the type of question used in the RAG pipeline. Within the context of synthetic hallucination generation, different types of questions were used to attempt to encourage hallucinations of different types. When examining the accuracy and F1 scores of base models and a fine-tuned model on these different types, we do not see significant differences in the performance across each type. The only exception to this is the higher accuracy on the Default Question type value.


F1 Score by Question Type
Average F1 Score
1.0
0.8
0.6
0.4
0.2
0.0
Errors, contradictions, or unsolvable questions
Multimodal content
Default question type
Advanced logical reasoning
Out-of-scope information
Other common hallucinated questions
Dataset Tag
Model
Qwen2-1.5B-Instruct
gpt-4o-mini
gpt-3.5-turbo
Qwen Finetune V1
claude-3-5-haiku-20241022
Qwen Finetune v2
claude-3-5-sonnet-20241022
gpt-4o
GPT-4o Mini Finetune


Accuracy by Question Type
Average Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Errors, contradictions, or unsolvable questions
Multimodal content
Advanced logical reasoning
Out-of-scope information
Other common hallucinated questions
Default question type
Dataset Tag
Model
Qwen2-1.5B-Instruct
gpt-4o-mini
gpt-3.5-turbo
Qwen Finetune V1
claude-3-5-haiku-20241022
Qwen Finetune v2
claude-3-5-sonnet-20241022
gpt-4o
GPT-4o Mini Finetune


F1 Score and Accuracy by Question Type (Aggregated across all models)
Average Value
0.8
0.6
0.4
0.2
0.0
Errors, contradictions, or unsolvable questions
Multimodal content
Default question type
Advanced logical reasoning
Out-of-scope information
Other common hallucinated questions
Dataset Tag
Metric
F1 Score
Accuracy

## Model performance of Base and FT Models by Hallucination Type

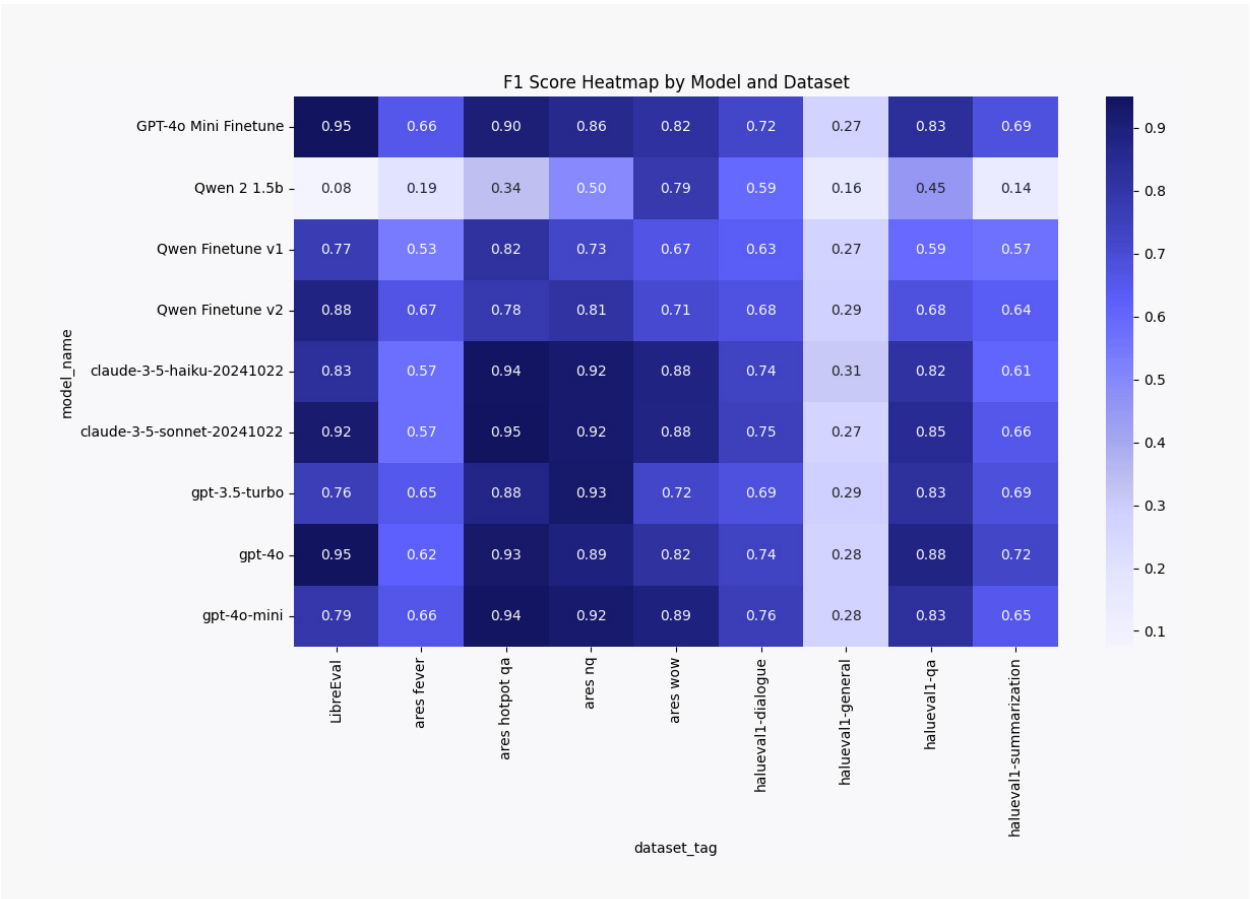


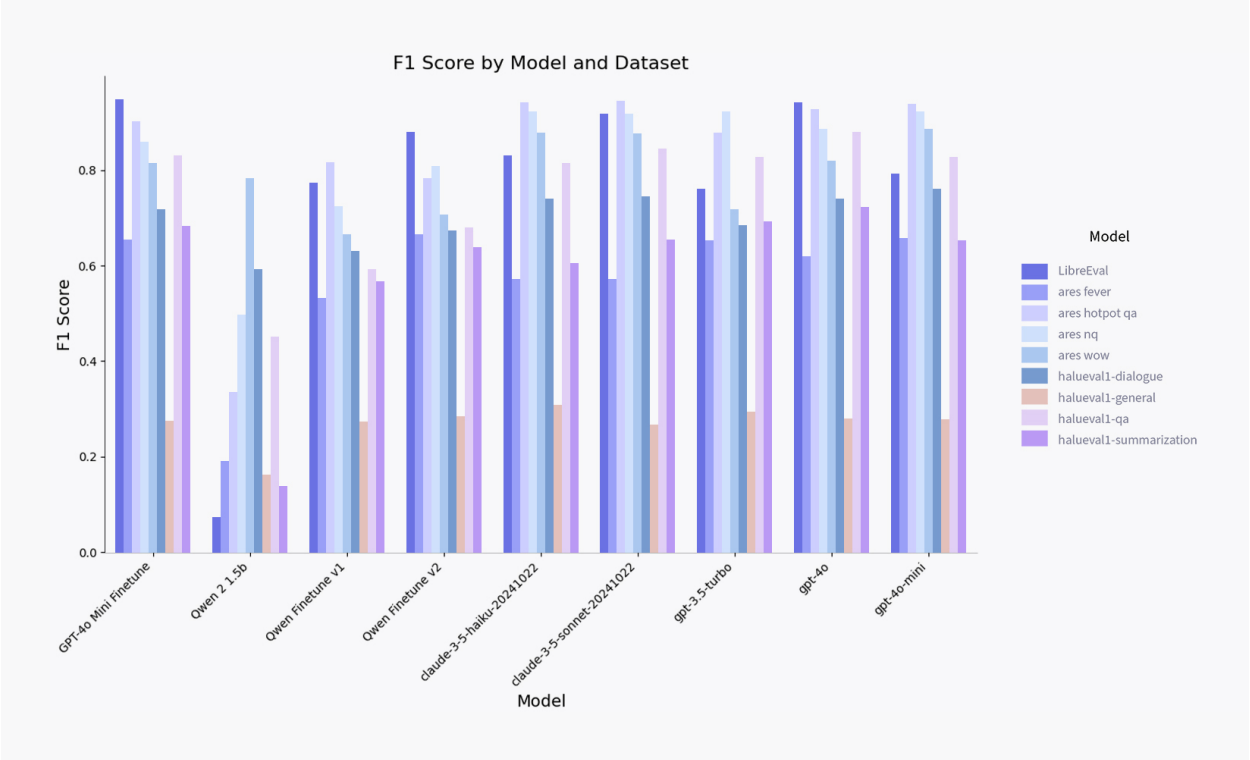

Both finetuned and base models showed the largest performance drop on Incompleteness hallucinations, where some of the information in the response text or text itself was missing. Aside from this, no single category of hallucinations appeared uniquely difficult for the judge models. GPT-4o-mini showed the highest variance in results, with a range of 0.3268.

## Finetuned Model Evaluation Performance Against Outside Datasets

Finetuned Qwen2 showed performance improvements over its base model. Qwen2 1.5b Instruct starts off with poor performance across the board, but responds well to fine-tuning, in some cases jumping over 80 points or a 10x increase. The only exception to this is on the Ares WoW dataset, where base Qwen2 outperforms its fine-tuned counterparts.


F1 Score by Dataset and Model
F1 Score
0.0
0.2
0.4
0.6
0.8
LibreEval
ares fever
ares hotpot qa
ares nq
ares wow
halueval1-dialogue
halueval1-dialogue
halueval1-qa
halueval1-summarization
Dataset
Model
Qwen2 1.5B
gpt-3.5-turbo
Qwen Finetune V1
gpt-4o-mini
claude-3-5-haiku-20241022
Qwen Finetune v2
claude-3-5-sonnet-20241022
gpt-4o
GPT-4o Mini Finetune


F1 Score Heatmap by Model and Dataset
model_name
GPT-4o Mini Finetune 0.95 0.66 0.90 0.86 0.82 0.72 0.27 0.83 0.69
Qwen 2 1.5b 0.08 0.19 0.34 0.50 0.79 0.59 0.16 0.45 0.14
Qwen Finetune v1 0.77 0.53 0.82 0.73 0.67 0.63 0.27 0.59 0.57
Qwen Finetune v2 0.88 0.67 0.78 0.81 0.71 0.68 0.29 0.68 0.64
claude-3-5-haiku-20241022 0.83 0.57 0.94 0.92 0.88 0.74 0.31 0.82 0.61
claude-3-5-sonnet-20241022 0.92 0.57 0.95 0.92 0.88 0.75 0.27 0.85 0.66
gpt-3.5-turbo 0.76 0.65 0.88 0.93 0.72 0.69 0.29 0.83 0.69
gpt-4o 0.95 0.62 0.93 0.89 0.82 0.74 0.28 0.88 0.72
gpt-4o-mini 0.79 0.66 0.94 0.92 0.89 0.76 0.28 0.83 0.65
LibreEval
ares fever
ares hotpot qa
ares nq
ares wow
halueval1-dialogue
halueval1-general
halueval1-qa
halueval1-summarization
dataset_tag
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

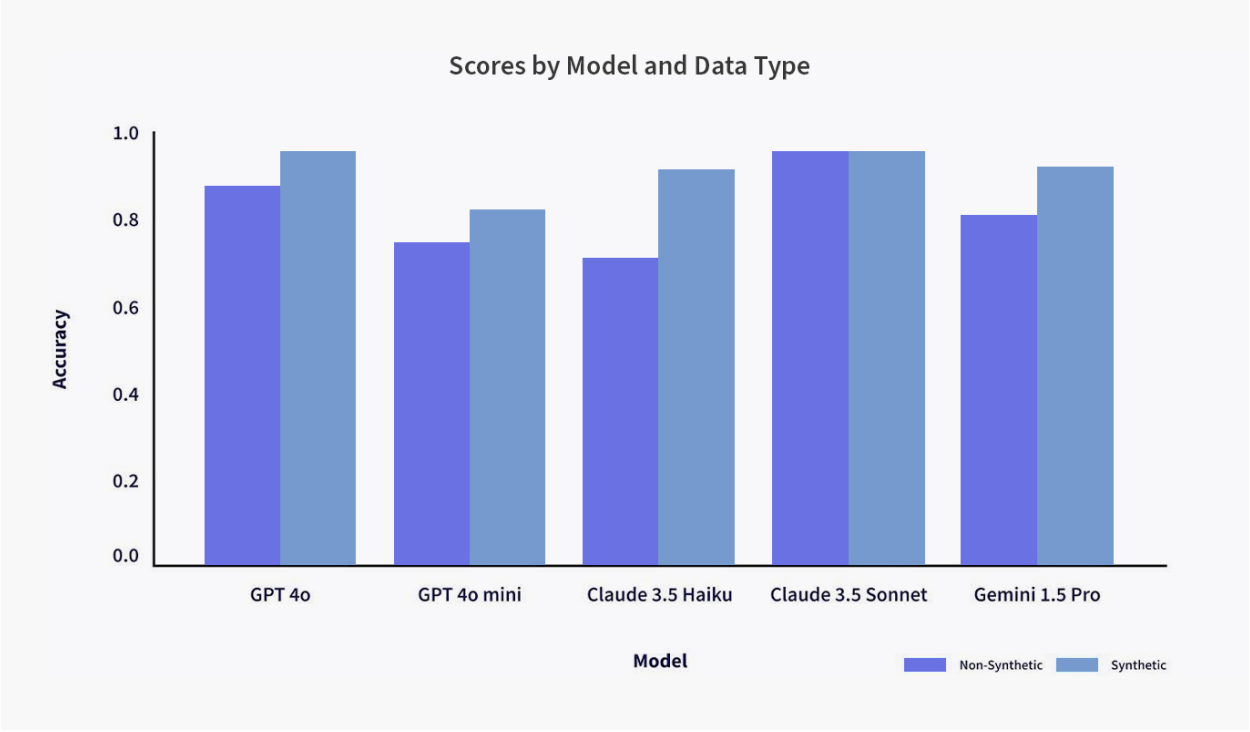



Interestingly, GPT-4o-mini often outperformed GPT-4o across five of the nine datasets, suggesting that even before fine-tuning, 4o-mini makes a strong candidate for hallucination detection.

## Ablations

### Base Model Evaluation Comparisons

Base models evaluated with the english test split performed better on synthetic data when base models were evaluated using Non-synthetic and Synthetic datasets. The GPT-4o and Claude 3.5 Sonnet models had the highest performance, and also were models used to generate the user input questions, and the responses for the synthetic data, which may have given them an advantage when evaluating hallucinations.




# Discussion

## Dataset Implications

The LibreEval1.0 platform and dataset effectively assessed hallucination evaluation performance across datasets segmented by knowledge domain, prompted question type, realized hallucination type, and synthetic versus non-synthetic data. This evaluation workflow helps identify weaknesses in existing datasets by leveraging the LibreEval platform to evaluate categorical performance for models tuned on data generated by the platform. Additionally, it guides future data generation efforts, ensuring well-rounded datasets that accurately detect hallucinations based on specific business needs. By leveraging LibreEval to scrape domain-specific web data and generate diverse hallucinations, we ensure that the dataset reflects real-world RAG applications and systematically addresses gaps in existing hallucination evaluation benchmarks.

The human hallucination labelling results demonstrate that the LLM Council of Judges serves as a highly reliable labeling mechanism. The decreased agreement between human annotators and the Council of Judges in the blind second test (81% vs. 96% in the first test) illustrates a more accurate comparison between the two approaches, as neither the human nor LLM judges have an existing label to compare against.

The observed patterns of disagreement provide important insights into the challenges of hallucination classification. The fact that disagreements were over four times more likely to occur in non-synthetic data suggests that real-world data suggests there are semantic complexities not present in synthetic examples. Moreover, the concentration of disagreements in Incompleteness Hallucinations and Relation Error Hallucinations indicates that these types remain particularly difficult for both human and LLM-based evaluation. This finding aligns with prior studies that highlight the challenge of identifying partial truth and relational inaccuracies within generative AI outputs.

Challenges still remain for creating datasets with ground-truth labels for spanning multiple categories of hallucination types. Our efforts used LLMs to generate different types of hallucinations and a council of LLMs to detect and classify these hallucinations. This approach was challenging because many generative models are specifically trained to avoid hallucinations in their outputs. When hallucination types were encouraged our council of LLM Judges found that no hallucinations were realized for a portion of the responses, specifically for encouraged hallucination types of Incompleteness Hallucination (87.53%), Outdated Information Hallucination (48.62%), Overclaim Hallucination (46.42%), and Unverifiable Information Hallucination (27.76%). Additionally, the use of a council of LLM judges still relies on the hallucination detection strength of the LLM models used.

## Models

Our study provides a direct comparison of proprietary and open-source models in hallucination evaluation tasks. While models such as GPT-4o and Claude-3.5-Sonnet excel in precision, recall scores vary significantly across different hallucination types. The gap between proprietary and fine-tuned open-source models is narrowing, particularly in applications that prioritize cost-efficiency and accessibility. This reinforces the value of continued investment in open-source fine-tuning efforts, as organizations can achieve strong performance without relying exclusively on high-cost proprietary APIs.

Our efforts show that on the HaluEval1.0 dataset, a small parameter and open-source model like Qwen2-1.5B-Instruct can be finetuned to achieve F1 performance slightly below a GPT-4o-mini base model performance (7% difference in English FT, 8% difference in multilingual FT). In largescale applications where the costs to use a proprietary model like GPT-4o-mini would be prohibitive, Qwen2-1.5B-Instruct offers an affordable self-hosted alternative with minimal overhead. We have open-sourced our finetunes for public use.

One limitation of our study is that our dataset contains question data that includes exact matches or semantic duplicates, which were distributed across our Train/Test/

Validation splits. This made our test datasets less effective for evaluating model performance of finetuned models, and led to our use of the HaluEval1.0 dataset to evaluate performance. A future improvement to the LibreEval1.0 platform should include methods to ensure semantically different questions generated during the data generation process.

## Future Work

While our study advances hallucination evaluation, several challenges remain. First, hallucination type classification remains an ambiguous task, with certain categories exhibiting low inter-judge agreement. Future work could explore refining LLM judge prompts or incorporating human oversight to improve consistency. Additionally, our dataset primarily focuses on English-language hallucinations, and while we have open-sourced a multilingual dataset, expanding fine-tuning efforts to more diverse languages remains a key area for further research. This is reflected in lack of hallucination detection in popular benchmarks today such as HaluEval1.0, HaluEval2.0, ARES, etc.

Additionally, our approach to combining human and LLM ensemble hallucination labels is relatively naive. Future efforts could explore increasing data trustworthiness by focusing human labeling on cases where the ensemble was not unanimous in its decision. Alternatively, humans could be incorporated as individual votes in a joint human-LLM ensemble. Either of these approaches could yield further interesting conclusions on human vs LLM judges.

Our study also focused entirely on a single method of hallucination detection through LLM as a Judge. Other techniques, like LMUnit (https://arxiv.org/abs/2412.13091) or Agent as a Judge (https://arxiv.org/abs/2410.10934)can produce more robust analysis of RAG systems by incorporating metrics beyond faithfulness. Future work could substitute alternative evaluation techniques in where LLM-as-a-Judge hallucination evaluation has been used in this study.

Lastly, real-world hallucination evaluation requires robustness against adversarial inputs. Our dataset was constructed from structured web data, but future efforts could explore dynamic evaluation frameworks that test models against user-generated adversarial samples. This would ensure that hallucination detection systems remain resilient to unpredictable and evolving challenges in deployed applications.

# Conclusion

The LibreEval1.0 platform and dataset represent a significant step forward in hallucination evaluation research. By systematically generating diverse hallucination types and benchmarking multiple models, we have provided insights into the strengths and limitations of current hallucination detection approaches. Our findings demonstrate that fine-tuned open-source models can serve as viable alternatives to proprietary solutions, particularly for cost-sensitive deployments. Future work should focus on refining hallucination categorization, improving multilingual evaluation, and enhancing model robustness through adversarial testing. By continuing to expand and refine LibreEval, we aim to contribute to more reliable and trustworthy AI-driven information retrieval systems.

# Appendix

## Data Generation Outcomes

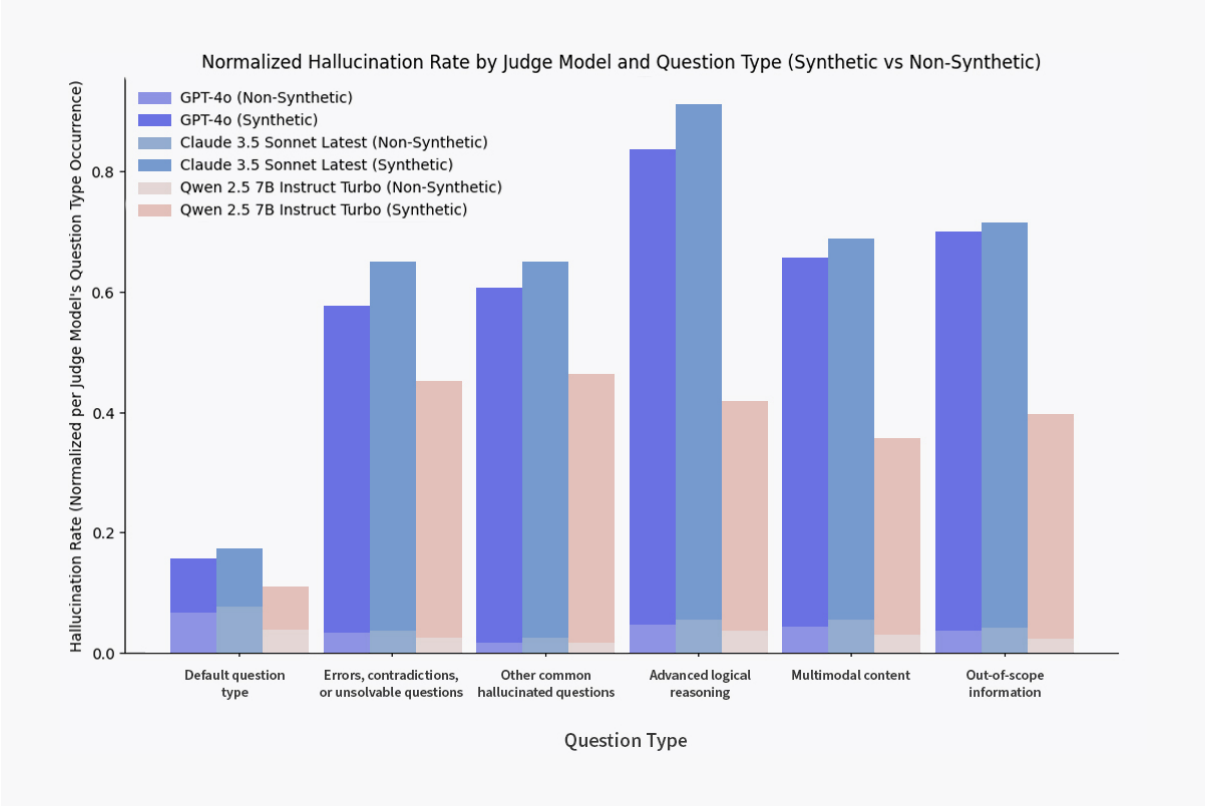
### Hallucination Labelling

#### Realized Hallucination Type by Prompted Question

A hallucination rate was calculated for each prompted question type to answer the question "For each model, how often does it hallucinate on a given question type, relative to how frequently it sees that question type? This was calculated using the following formula:

$$\textit{Hallucination Rate} = \frac{\textit{Hallucinated Count (by model, question type, synthetic status)}}{\textit{Total Occurences (by model, question type)}}$$

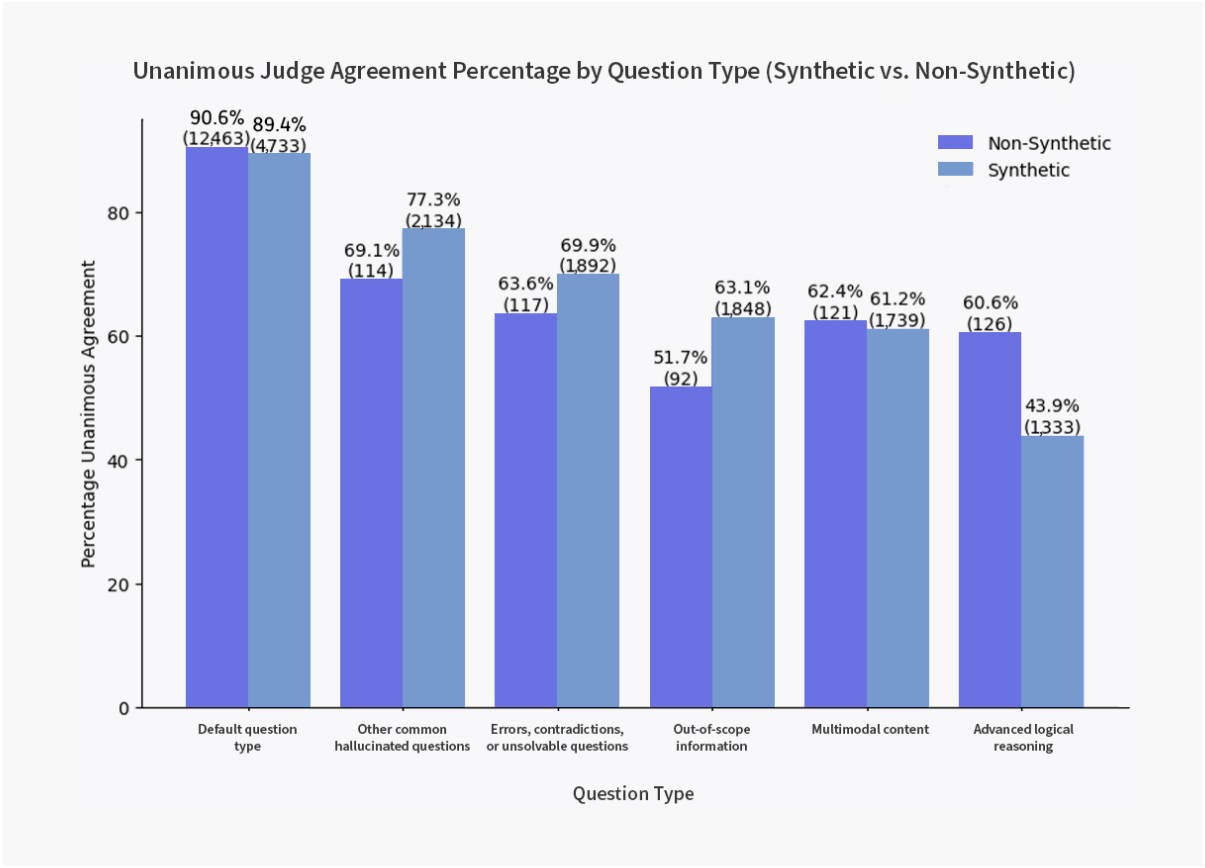
Advanced Logical Reasoning remains the most hallucination-prone category across all models, with synthetic questions exhibiting a higher hallucination rate than non-synthetic ones, suggesting increased difficulty in these tasks. Multimodal Content and Out-of-Scope Information also show consistently high hallucination rates, particularly for GPT-4o and Claude 3.5 Sonnet. Errors, Contradictions, or Unsolvable Questions and Other Common Hallucinated Questions exhibit moderate hallucination rates, with synthetic data amplifying hallucination tendencies. Default Question Type maintains the lowest hallucination rate across all models, with minimal differences between synthetic and non-synthetic cases, indicating strong model performance on standard question types without excessive hallucinations.




### Judge Hallucination Agreement Based Upon Prompted Question

The "Unanimous Judge Agreement Percentage by Question Type (Synthetic vs. Non-Synthetic)" graph shows that certain question types had lower agreement among judges. Advanced logical reasoning had the lowest overall agreement, with only 43.9% in synthetic data and 60.6% in non-synthetic data, indicating that these questions remain the most challenging for judges to classify consistently. Multimodal content also exhibited relatively low agreement, with 61.2% for non-synthetic data and 62.4% for synthetic data, suggesting variability in judgments. In contrast, default question types had the highest agreement, with 90.6% in non-synthetic and 89.4% in synthetic data, showing minimal ambiguity.

The differences in unanimous agreement between synthetic and non-synthetic data remain within an acceptable range (<10%) for most question types, including default questions, errors/contradictions, and out-of-scope information. However, advanced logical reasoning showed a notable 16.7% difference, reinforcing that synthetic data introduces greater inconsistency in this category. Similarly, multimodal content had a smaller but meaningful gap of 1.2%, indicating some difficulty in synthetic data handling non-text-based inputs. Overall, while agreement levels remain similar across most categories, complex reasoning and multimodal questions show the highest judgment variability, particularly in synthetic data.

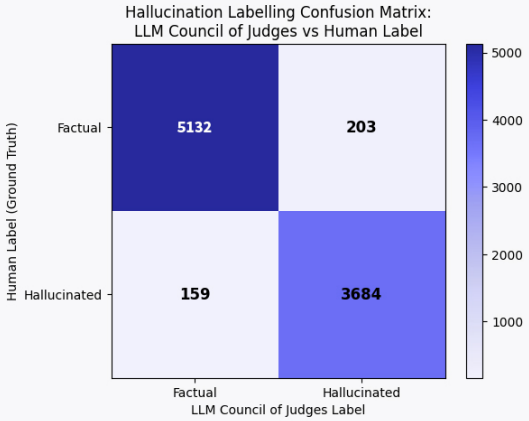


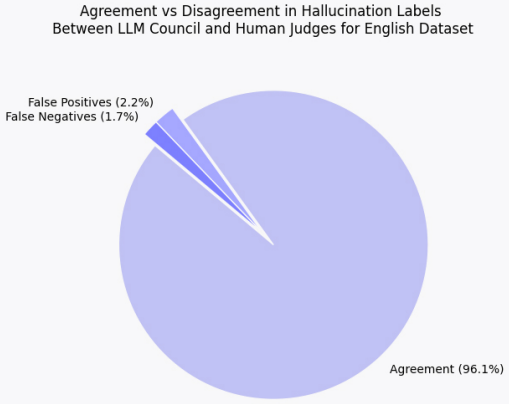

**Hallucination Labels from Human and LLM Council of Judges Labellers in the English dataset**

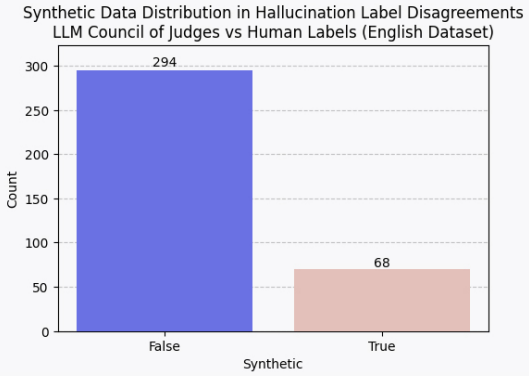
Our evaluation of hallucination labeling consistency between human annotators and the LLM Council of Judges yielded insightful findings regarding agreement rates and reliability. In our initial test of 100 samples, human annotators and the LLM Council of Judges agreed 96% of the time. However, manual review by the authors determined that human labels were correct 93% of the time, while the LLM Council of Judges' labels were correct 94% of the time. This slight edge in accuracy for the Council of Judges suggested its effectiveness as a primary source of labeling.

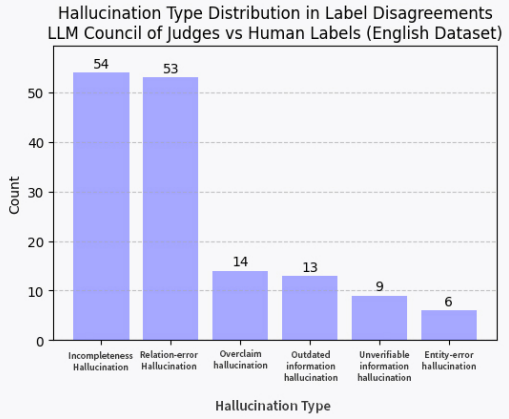
A subsequent blind evaluation of 100 new samples revealed a decrease in agreement between human annotators and the LLM Council of Judges, falling to 81%. Despite this drop, manual review found that the human annotators were correct in 85% of cases, whereas the LLM Council of Judges demonstrated a higher correctness rate at 92%. This reinforced the initial finding that the LLM Council of Judges provided highly reliable hallucination labeling. As a result, we prioritized the Council of Judges' majority labels in training and evaluation while incorporating human labels in 10% of cases for comparative analysis.

Further analysis of disagreements between the LLM Council of Judges and human annotators revealed key patterns. Overall, the two sources agreed in 96.1% of cases, with false negatives accounting for 1.7% and false positives for 2.2%. Notably, disagreements were over four times more likely to occur in non-synthetic data than in synthetic data, suggesting that human annotators and the LLM Council of Judges exhibited greater consistency in labeling synthetic content. Additionally, the most common sources of disagreement were related to Incompleteness Hallucinations and Relation Error Hallucinations, regardless of whether the data was synthetic or non-synthetic.



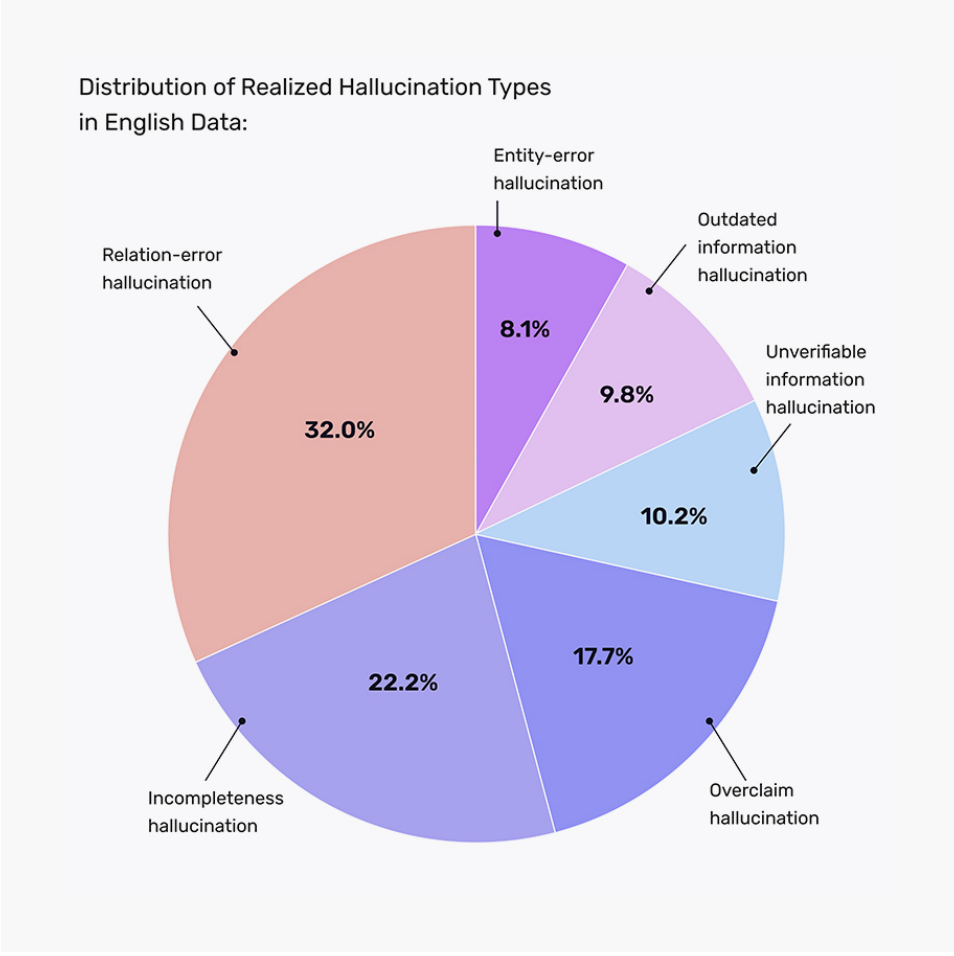



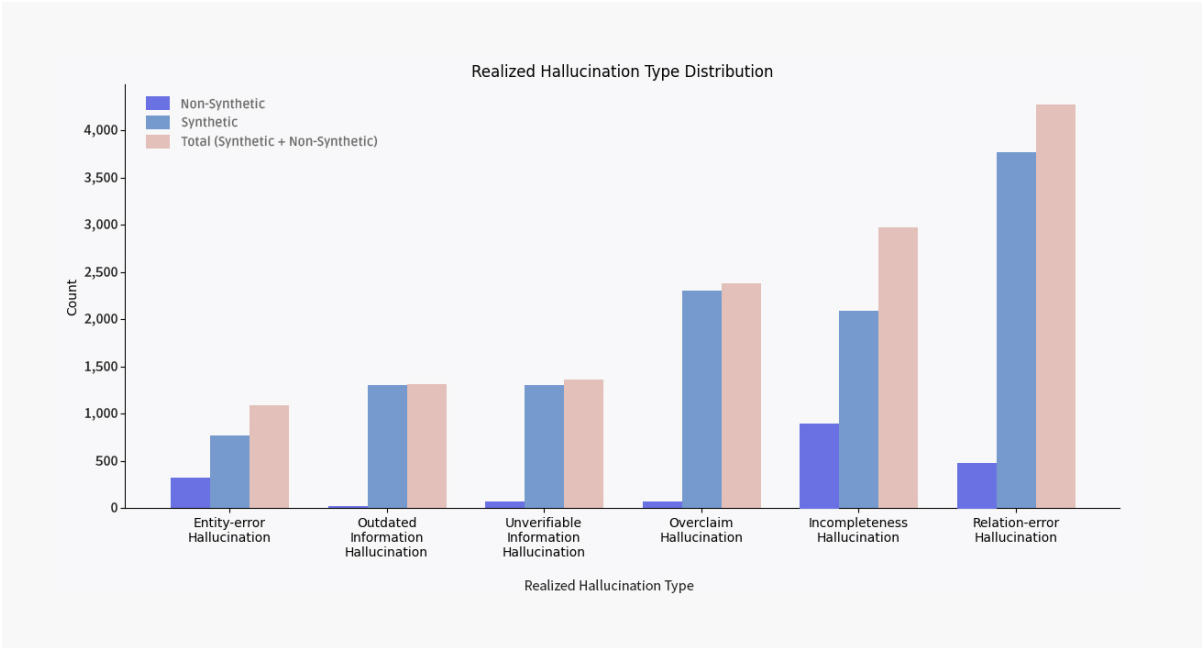
### Hallucination Type Labelling

The hallucination type distribution captures both frequent and high-impact failure modes, making it well-suited for training hallucination detection models. Relation-error hallucinations (32.0%) are the most prevalent, highlighting issues with incorrect entity relationships, followed by incompleteness hallucinations (22.2%), which reflect missing critical details in responses. Overclaim hallucinations (17.7%) represent exaggerated or misrepresented facts, while unverifiable information (10.2%) and outdated information (9.8%) address risks in factual reliability. Entity-error hallucinations (8.1%), though less common, are crucial for ensuring precision in high-stakes domains such as law, medicine, and finance.



### Categories of Hallucinated Response Among Synthetic and Non-Synthetic Data

In non-synthetic data, “Incompleteness” and “Relation-error Hallucinations” dominate, while “Outdated Information,” “Unverifiable Information,” and “Overclaim Hallucinations” are almost nonexistent. Synthetic data amplifies all types, especially “Relation-error” and “Incompleteness Hallucinations,” making them the most frequent overall. The scarcity of “Outdated Information Hallucinations” in non-synthetic data suggests that response-generating and hallucination-judging models share similar training data, reducing such errors naturally.

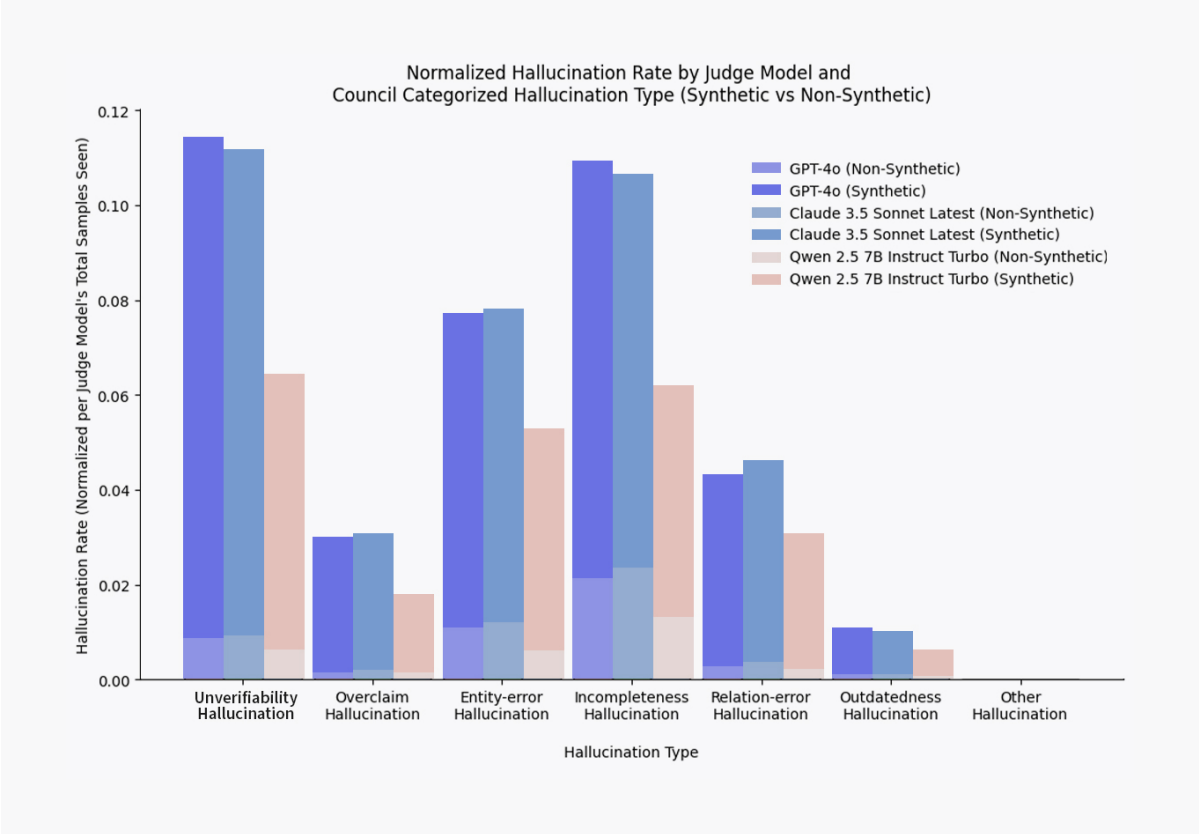



### Judge Model Hallucination Type Rates of Detection

In order to answer the question: “"For each model, how frequently does it detect different types of hallucinations, normalized by the total number of samples it judged? How does this compare between synthetic and non-synthetic data?" a hallucination rate was calculated for each hallucination type. This was calculated using the following formula:

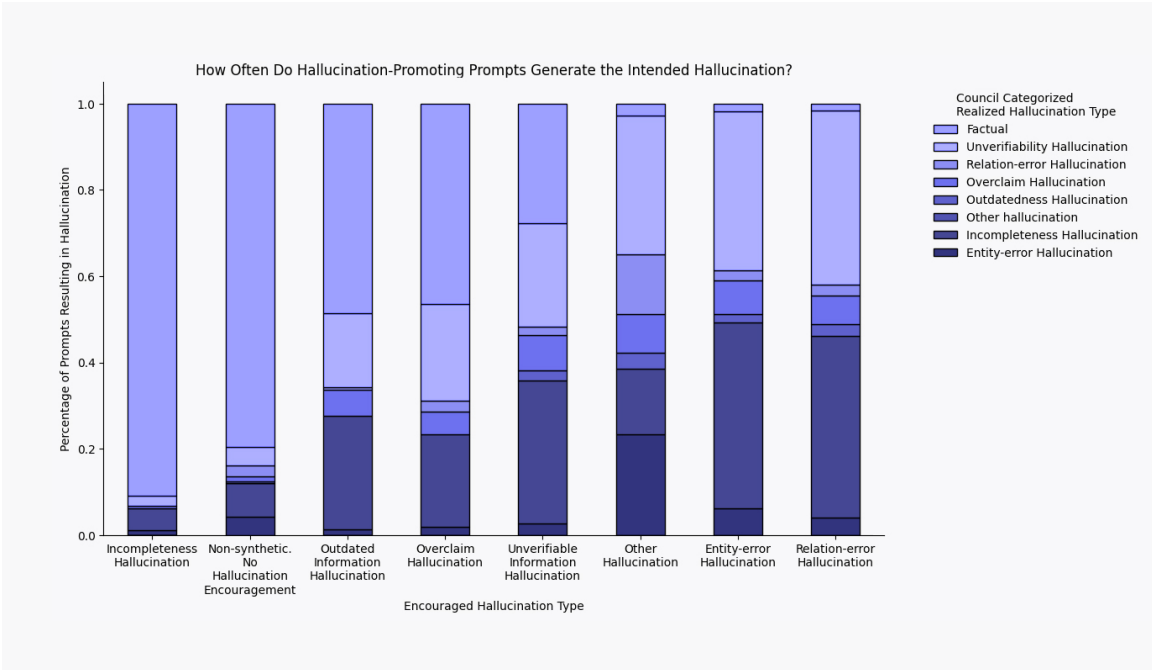
$$\text{Hallucination Rate} = \frac{\text{Hallucinated Count (by model, hallucination type, synthetic status)}}{\text{Total Samples Seen by Model}}$$

> Relation-Error Hallucination has the highest hallucination rate across all models, followed closely by Incompleteness Hallucination, indicating that models struggle most with maintaining factual consistency and completeness. Entity-Error Hallucinations and Outdated Information Hallucinations have the lowest rates, suggesting that models generally avoid fabricating named entities or outdated details. Across all hallucination types, synthetic data consistently leads to higher hallucination rates, reinforcing the idea that synthetic cases are more challenging for models, likely due to adversarial design or increased complexity in synthetic prompts.



### Realization of Encouraged Hallucination Type

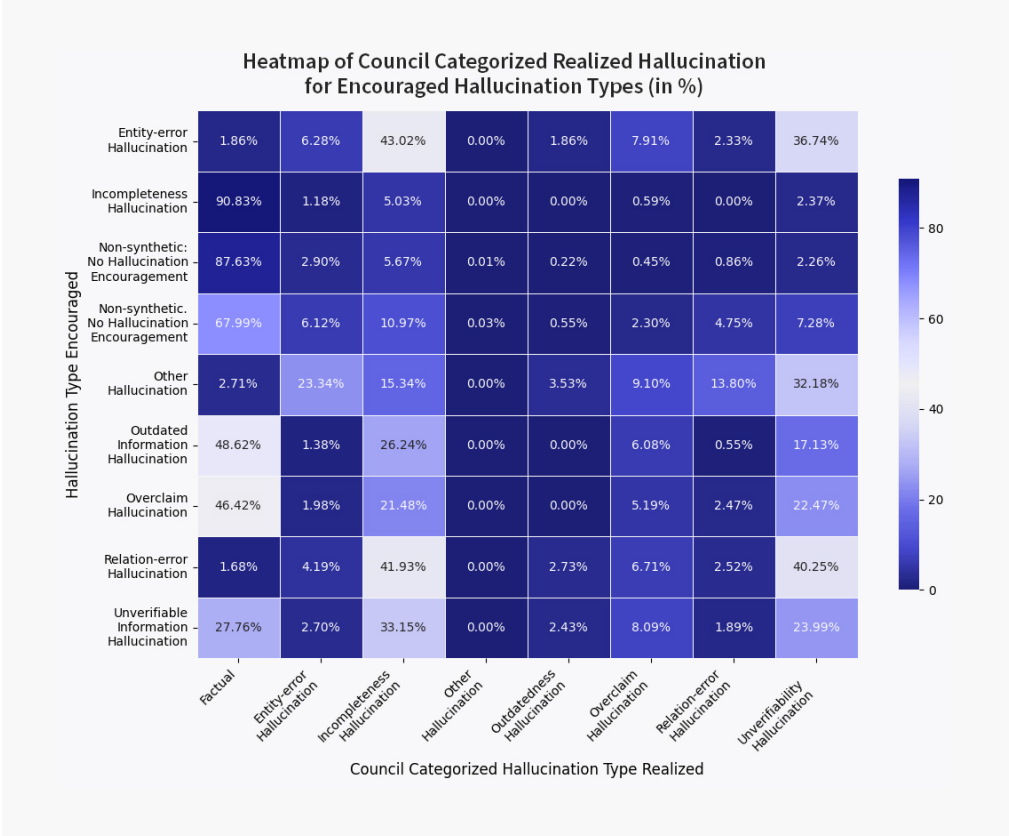
To assess whether prompts reliably induce the intended hallucination type, we compare encouraged vs. realized hallucinations using LLM-based evaluation. "Factual" cases indicate responses judged as factually correct, meaning no hallucination was detected. "Non-Synthetic: No Hallucination Encouragement" serves as a baseline, showing how often hallucinations arise naturally without explicit prompting. While some hallucination types, such as Entity-Error and Relation-Error Hallucinations, closely align with their intended categories, others—like Overclaim and Unverifiable Information Hallucinations—often result in a mix of unintended hallucinations. This suggests that while targeted prompts can guide hallucination type generation, responses remain variable, with some categories being more prone to cross-type hallucination errors.




The data suggests that when no hallucination encouragement is provided, a large majority (87.53%) of responses are classified as factual, meaning that the LLM judge did not detect hallucination. However, hallucinations still emerge naturally, with 5.96% of responses classified as Incompleteness Hallucinations and 3.33% as Relation-Error Hallucinations, highlighting that even neutral prompts can lead to systematic factual distortions. Interestingly, when Incompleteness Hallucinations were explicitly encouraged, the factual rate remained high (90.83%), suggesting that the model largely resists producing incomplete responses even when prompted to do so.

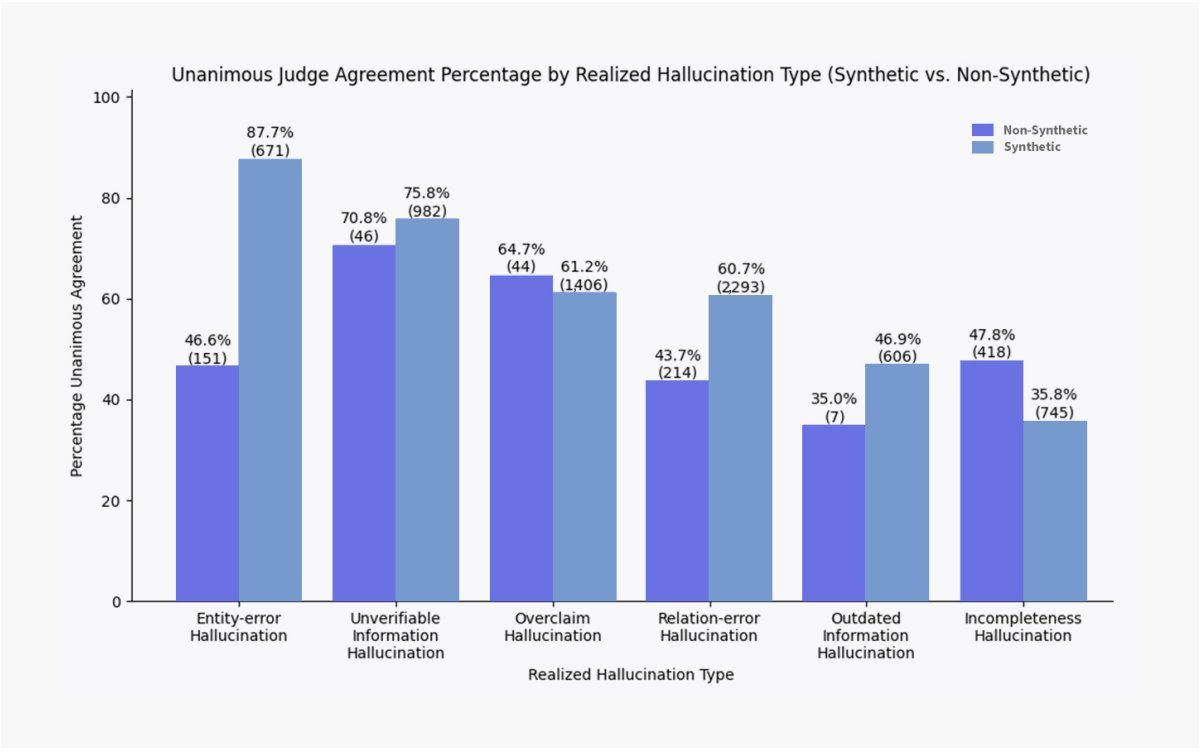
When specific hallucinations are encouraged, the model's responses generally align with the intended hallucination category, though cross-category leakage persists. Entity-Error and Relation-Error Hallucinations frequently overlap, with 31.45% of Relation-Error Encouragement cases resulting in Entity-Error Hallucinations, suggesting that these hallucination types are difficult for LLM hallucination evaluators to distinguish. Additionally, Overclaim and Outdated Information Hallucinations maintain moderate factual rates (46.42% and 48.62%, respectively), indicating that while the model is susceptible to these hallucinations, it does not always comply with the hallucination prompt. Notably, Unverifiable Information Encouragement resulted in a high rate of mixed hallucination types, with 12.94% classified as Entity-Error, 7.01% as Relation-Error, and 18.06% as Unverifiable Information itself, highlighting the difficulty in reliably inducing one specific hallucination type.

Since hallucination detection itself is performed by an LLM, these results should be interpreted as an approximation of hallucination trends rather than absolute ground truth. However, the data suggests that certain hallucination types, such as Entity-Error and Relation-Error Hallucinations, exhibit significant ambiguity in classification, while others, like Outdated Information and Overclaim Hallucinations, show partial resistance from the model, resulting in a mix of factual and hallucinated responses.



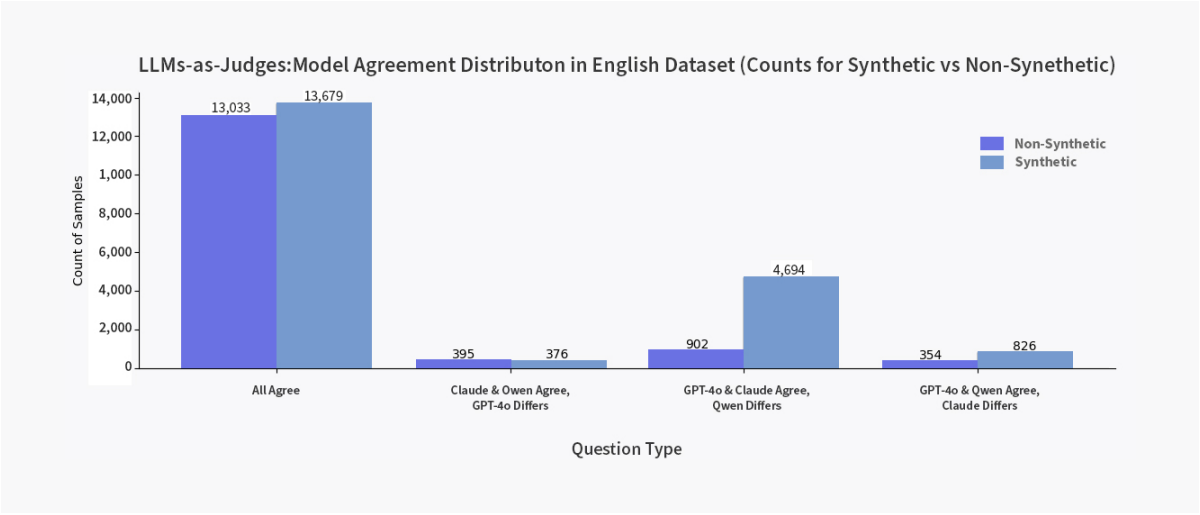
The "Unanimous Judge Agreement Percentage by Realized Hallucination Type (Synthetic vs. Non-Synthetic)" graph shows that certain hallucination types had significantly lower agreement among judges. Incompleteness hallucinations had the lowest unanimous agreement, with 35.8% in synthetic data and 47.8% in non-synthetic data, suggesting high ambiguity in classification. Outdated information hallucinations also exhibited lower agreement (35.0% for non-synthetic vs. 46.9% for synthetic), indicating variability in how these errors are judged.

Most hallucination types show a moderate difference in agreement between synthetic and non-synthetic data, including overclaim (75.8% vs. 70.8%) and relation-error hallucinations (61.2% vs. 64.7%). However, entity-error hallucinations had a substantial gap, with 87.7% agreement in synthetic data versus only 45.5% in non-synthetic data, suggesting that synthetic data may introduce more consistency in labeling. Overall, while most hallucination types have comparable agreement levels, entity-error hallucinations exhibit a significant difference, potentially reflecting differences in how these errors manifest in synthetic vs. real-world cases. For analysis of synthetic and non-synthetic judge agreement across hallucination types please refer to the Judge Agreement Across Hallucination Types section of the appendix.



### Council of Judges Agreement Across Synthetic and Non-Synthetic Datasets for Hallucination Labels

A majority of hallucination labels were unanimously agreed upon by all judges, with higher consensus amongst synthetic data. The most disagreement occurred where the label generated by the Qwen model differed from GPT-4o and Claude.




### Token Distributions

The following graphs present the token distributions for key input components used in hallucination evaluation: input questions, reference context, output responses, and combined input + reference tokens. These distributions provide insight into the length variability across synthetic and non-synthetic data, aiding in the selection of models with appropriate token limitations for downstream applications. We have also provided token counts to offer context for those selecting models with a maximum token limit and to give ballpark estimates for cost considerations when fine-tuning on APIs that bill by token count.

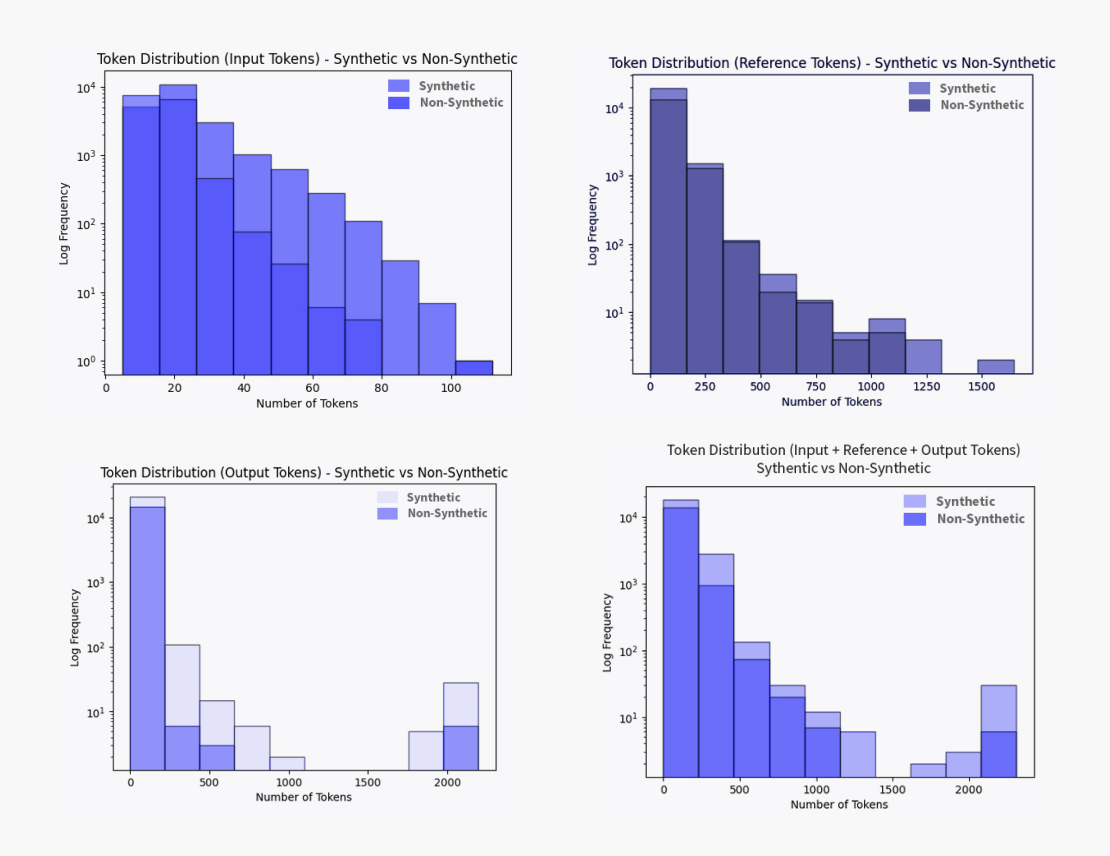
#### Token Distributions for Hallucination Evaluation Inputs

- **Input Tokens:** The input question token distribution shows that most input questions are under 60 tokens, with a maximum length below 100 tokens. Synthetic data exhibits a higher proportion of longer questions, suggesting that synthetically generated queries tend to be more detailed or complex compared to non-synthetic ones.
- **Reference Tokens:** The reference text token distribution indicates that most reference texts are under 500 tokens, with a maximum size of approximately 1500

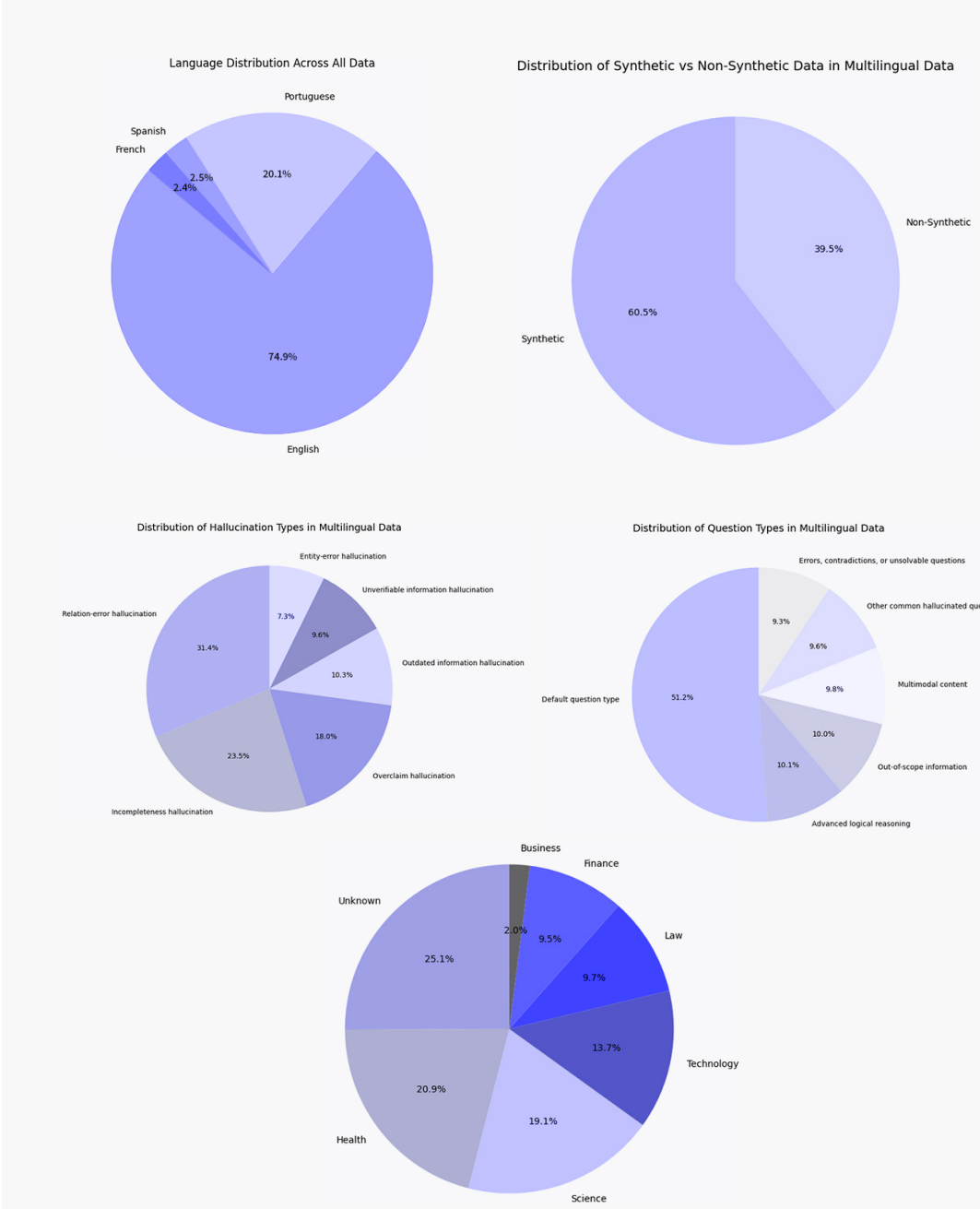
tokens. Both synthetic and non-synthetic references follow a similar pattern, though synthetic references display slightly fewer extremely long samples, potentially due to controlled dataset generation.

- **Output Tokens:** The output response token distribution reveals that most outputs contain fewer than 250 tokens, though some extend beyond 2000 tokens. Synthetic responses tend to be longer, with higher token frequencies at the upper length range, suggesting that synthetic responses are more verbose and detailed than their non-synthetic counterparts.
- **Total Tokens (Excluding Prompt):** The input + reference + output token distribution highlights that most samples fall below 500 tokens, aligning with common LLM context window limits. However, longer token sequences (up to 2000+ tokens) are more prevalent in synthetic data, making this dataset valuable for training models that require exposure to both standard-length and extended text sequences.

### Token Counts Distributions for English Dataset



## Multilingual Dataset Distributions

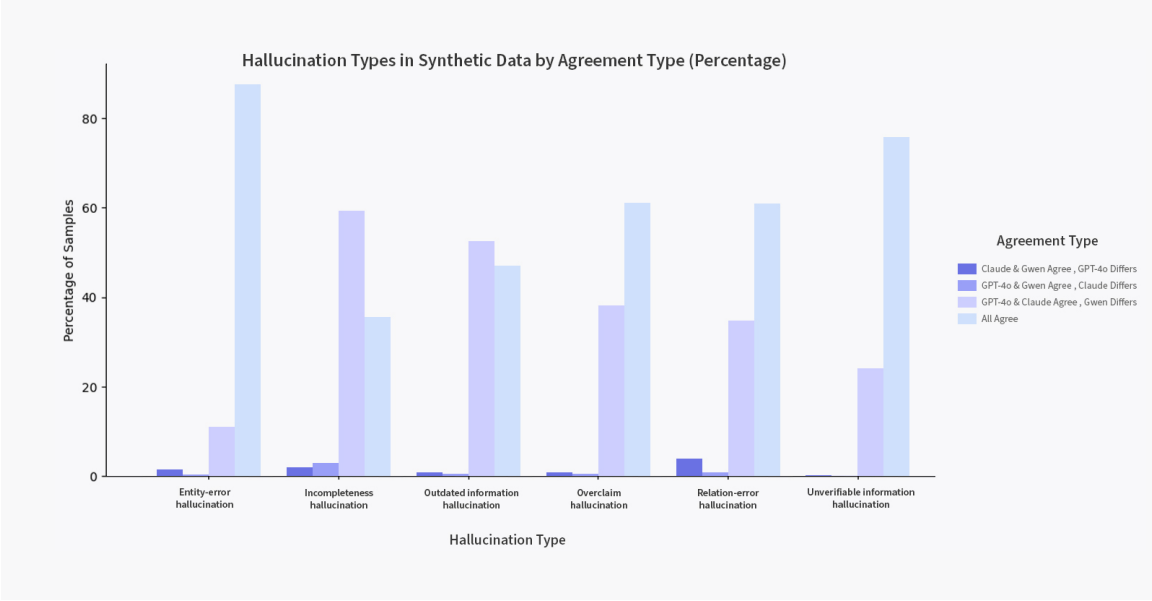


**European Multilingual Dataset Sample Count Statistics (EN-ES-FR-PT) for All Data**

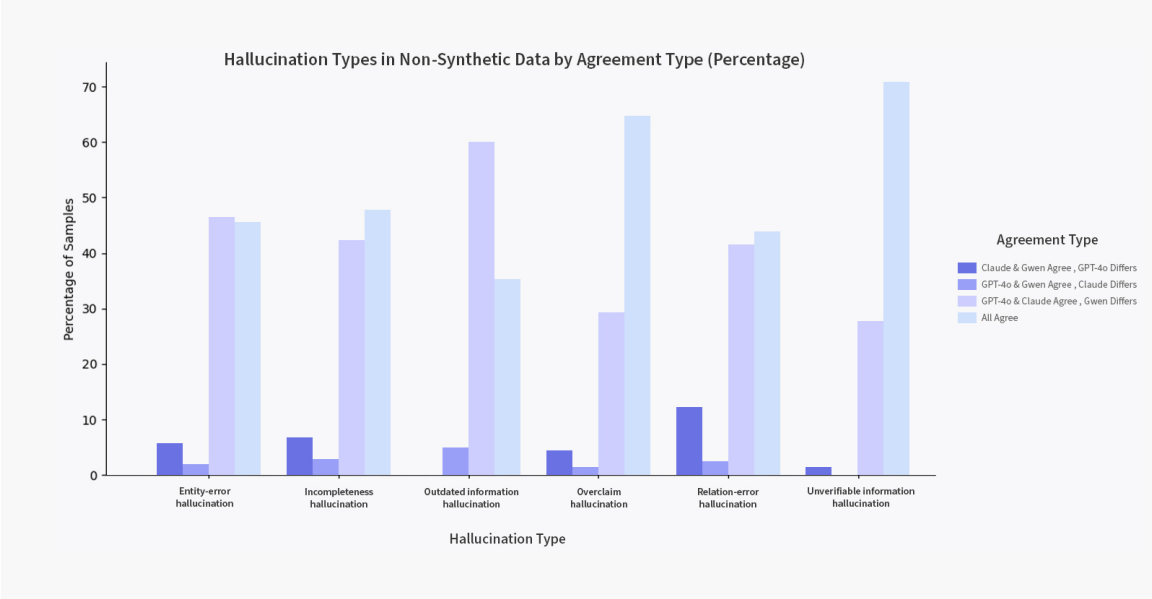
| Synthetic | Language | Factual | Hallucinated | Total |
|---|---|---|---|---|
| Non-Synthetic | English | 12,853 | 1,831 | 14,684 |
| | Spanish | 316 | 78 | 394 |
| | French | 319 | 63 | 382 |
| | Portuguese | 2,768 | 639 | 3,407 |
| Synthetic | English | 8,612 | 12,479 | 21,091 |
| | Spanish | 404 | 402 | 806 |
| | French | 381 | 404 | 785 |
| | Portuguese | 2,962 | 3,232 | 6,194 |

### Judge Agreement Across Hallucination Types

Across synthetic data, most hallucination types showed >50% council consensus, except Incompleteness Hallucinations (35.8%) and Outdated Information Hallucinations (37.9%), which had the highest disagreement. Entity-Error (87.7%) and Unverifiable Information Hallucinations (75.8%) had the strongest agreement. Qwen was the most frequent outlier, often diverging from GPT-4o and Claude. Overclaim (61.2%) and Relation-Error Hallucinations (60.7%) had moderate consensus but still notable cross-model disagreement, indicating persistent ambiguity in these categories.




In non-synthetic data, all hallucination types had >50% council consensus, with Unverifiable Information (70.8%) and Overclaim Hallucinations (64.7%) showing the highest agreement. Outdated Information Hallucinations (35.0%) had the lowest consensus, making them the most ambiguous to evaluate. Qwen was the most frequent outlier, often diverging from GPT-4o and Claude. While most hallucination types had strong agreement, Outdated Information Hallucinations remain difficult to classify consistently across models.



# Prompts

## Hallucination Type Definitions and their Corresponding Generation Prompts

| Hallucination Type | Description | Prompt |
|---|---|---|
| Relation-error hallucination | Introducing incorrect relationships between entities, such as quantitative, temporal, or cause-effect relations. | Answer the following question by subtly introducing incorrect relationships between entities, such as quantitative, temporal, or cause-effect relations. Only respond with the anwer, no other text. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |
| Incompleteness hallucination | Providing a response that appears complete but omits some critical details or facts. | Answer the following question with a response that appears complete but omits some critical details or facts when listing or aggregating information.Only respond with the answer, no other text. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |
| Outdated information hallucination | Providing a response that subtly includes outdated information as if it were current. | Answer the following question by providing a response that subtly includes outdated information as if it were current. Only respond with the answer, no other text. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |
| Overclaim hallucination | Making a statement that subtly exaggerates or overclaims beyond the factual or logical scope. | Answer the following question with a statement that very subtly exaggerates or overclaims beyond the factual or logical scope of the given information. This should be a very subtle exaggeration. Only respond with the answer, no other text. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |
| Unverifiable information hallucination | Providing information that cannot be verified using commonly available sources, ensuring it sounds plausible but unverifiable. | Answer the following question with information that cannot be verified using commonly available sources, ensuring it sounds plausible but unverifiable. Only respond with the answer, no other text. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |
| Entity-error hallucination | Introducing incorrect entities, such as wrong names, dates, locations, or objects. | Answer the following question by subtly introducing incorrect entities, such as the wrong names, dates, locations, or objects, in a way that blends with the context. Make sure the answer is in {language}.<br><br>Context: {paragraph}<br><br>Question: {question}<br><br>Answer: |

## Question Type Generation Prompts

| Question Type | Description | Prompt |
|---|---|---|
| System Prompt | System prompt for generating questions. | You are a helpful assistant that generates questions based on given content. Only respond with the question, no other text. The questions should be in {language}. |
| Out-of-scope information | Seeking details not present in the model's training data, such as real-time or future information, asking for external links, or seeking highly specific, subjective, or personal information. | Generate a question that can be answered using the following paragraph. Make sure the question seeks information about events occurring in the future, references external websites or links, or asks for highly specific and subjective interpretations. Make sure the question is in {language} Only respond with the question, no other text. Do not introduce the question in any way.<br><br>Paragraph:<br><br>{paragraph} |
| Advanced logical reasoning | Challenging requests that surpass the model's capacity for logical reasoning and problem-solving, including intricate mathematical or programming problems. | Generate a question that can be answered using the following paragraph. Ensure the question requires advanced logical reasoning or solving an intricate mathematical or programming problem. Make sure the question is in {language}. Only respond with the question, no other text. Do not introduce the question in any way.<br><br>Paragraph:<br><br>{paragraph} |
| Multimodal content | Seeking output beyond text, such as images, sound, or videos, which is beyond the usual capabilities of language models primarily designed for text-based tasks. | Generate a question that can be answered using the following paragraph.Ensure the question asks for content beyond text, such as images, sounds, or videos. Make sure the question is in {language}. Only respond with the question, no other text. Do not introduce the question in any way.<br><br>Paragraph:<br><br>{paragraph} |
| Errors, contradictions, or unsolvable questions | Queries that contain errors within themselves, such as nonsensical strings, invalid or ambiguous inputs, unsolvable questions, or false statements. | Generate a question that can be answered using the following paragraph.Make sure the question includes errors, contradictions, or is otherwise invalid, ambiguous, or unsolvable. Make sure the question is in {language}. Only respond with the question, no other text. Do not introduce the question in any way.<br><br>Paragraph:<br><br>{paragraph} |
| Other common hallucinated questions | The query is easy to cause hallucination but is not covered in the above 5 types. | Generate a question that can be answered using the following paragraph. Ensure the question is structured in a way that commonly leads to hallucination. Make sure the question is in {language} Only respond with the question, no other text. Do not introduce the question in any way.<br><br>Paragraph:<br><br>{paragraph} |

## LabelBox Human Labeller Instruction - Test 1
(LLM judge responses provided to human labelers)

In this task, you will be presented with a query, a reference text and an answer. The answer is generated to the question based on the reference text. The answer may contain false information. You must use the reference text to determine if the answer to the question contains false information, if the answer is a hallucination of facts. Your objective is to determine whether the answer text contains factual information and is not a hallucination. A 'hallucination' refers to an answer that is not based on the reference text or assumes information that is not available in the reference text.

- "hallucinated" indicates that the answer provides factually inaccurate information to the query based on the reference text.

- "factual" indicates that the answer to the question is correct relative to the reference text, and does not contain made up information.

Please read the query and reference text carefully before determining your response.

We have also provided a label that's been generated by a set of AI judges, along with an explanation of their choice. **The judges' response may be correct or incorrect**.

Is the answer above factual or hallucinated based on the query and reference text?

## LabelBox Human Labeller Instruction - Test 2
(LLM judge responses not provided to human labelers)

In this task, you will be presented with a query, a reference text and an answer. The answer is generated to the question based on the reference text. The answer may contain false information. You must use the reference text to determine if the answer to the question contains false information, if the answer is a hallucination of facts. Your objective is to determine whether the answer text contains factual information and is not a hallucination. A 'hallucination' refers to an answer that is not based on the reference text or assumes information that is not available in the reference text.

- "hallucinated" indicates that the answer provides factually inaccurate information to the query based on the reference text.

- "factual" indicates that the answer to the question is correct relative to the reference text, and does not contain made up information.

Please read the query and reference text carefully before determining your response.

Is the answer above factual or hallucinated based on the query and reference text?