Confabulations from acl publications (cap): A dataset for scientific hallucination detection

# Confabulations from ACL Publications (CAP): A Dataset for Scientific Hallucination Detection

**Federica Gamba**[*] **Aman Sinha**[*] **Timothee Mickus**
**Raúl Vázquez** **Patanjali Bhamidipati** **Claudio Savelli**
**Ahana Chattopadhyay** **Laura Zanella** **Yash Kankanampati**
**Binesh Arakkal Remesh** **Aryan Chandramania** **Rohit Agarwal**
**Chuyuan Li** **Ioana Buhnila** **Radhika Mamidi**

[*]These authors have equal contribution.

Université de Lorraine, France; Charles University, Prague; Independent Researcher;
University of Helsinki, Finland; Université Sorbonne Paris Nord, France;
IIIT Hyderabad, India; Politecnico di Torino, Italy; UiT Tromso, Norway;
University of British Columbia, Vancouver, Canada Chosun University, South Korea

**Correspondence:** {gamba@ufal.mff.cuni.cz, aman.sinha@univ-lorraine.fr}

**Abstract**

We introduce the CAP (Confabulations from ACL Publications) dataset, a multilingual resource for studying hallucinations in large language models (LLMs) within scientific text generation. CAP focuses on the scientific domain, where hallucinations can distort factual knowledge, as they frequently do. In this domain, however, the presence of specialized terminology, statistical reasoning, and context-dependent interpretations further exacerbates these distortions, particularly given LLMs' lack of true comprehension, limited contextual understanding, and bias toward surface-level generalization. CAP operates in a cross-lingual setting covering five high-resource languages (English, French, Hindi, Italian, and Spanish) and four low-resource languages (Bengali, Gujarati, Malayalam, and Telugu). The dataset comprises 900 curated scientific questions and over 7,000 LLM-generated answers from 16 publicly available models, provided as question–answer pairs along with token sequences and corresponding logits. Each instance is annotated with a binary label indicating the presence of a scientific hallucination, denoted as a factuality error, and a fluency label, capturing issues in the linguistic quality or naturalness of the text. CAP is publicly released to facilitate advanced research on hallucination detection, multilingual evaluation of LLMs, and the development of more reliable scientific NLP systems.

**Keywords:** Hallucination detection, Multilingual NLP, Scientific text generation

Helsinki-NLP/shroom-cap
Helsinki-NLP/shroom_cap

## 1. Introduction

As the prevalence of large language model (LLM) technology grows, so do concerns about its reliability and trustworthiness. This state of affairs stems from the general ambivalence of these systems when it comes to the truthfulness of their outputs (Hicks et al., 2024; van Deemter, 2024; Perez et al., 2023) — these models grow more fluent, but they need not output sentences that are factually correct, which gives rise to fluent but factually false outputs, or 'hallucinations'. Remarkably, researchers interested in hallucinations often emphasize hallucinations as issues of factuality: Kalai et al. (2025) define hallucinations as "plausible falsehoods", whereas van Deemter (2024) provides a taxonomy based on logical contradictions. Such approaches commonly assume that LLMs consistently produce fluent and coherent text across languages. This assumption, however, does not necessarily hold in the context of *low-resource languages*.

In languages with limited representation in training corpora, LLMs frequently exhibit reduced fluency, degraded coherence, and a higher incidence of semantically or syntactically flawed outputs (Dargis et al., 2024). These limitations can result in hallucinations that go beyond factual inaccuracy, including structural errors, semantic drift, or outright nonsensical generations. The scarcity of high-quality datasets for such languages constrains both model training and evaluation, making it difficult to assess and ensure reliability. Together, these observations underscore the need for language technologies that account for linguistic diversity and for benchmark resources that can capture model behavior more accurately across underrepresented languages.

We argue that the mechanisms underlying hallucinatory behaviors in LLMs thus stand at the cross-



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2512–2524
11–16 May 2026. 

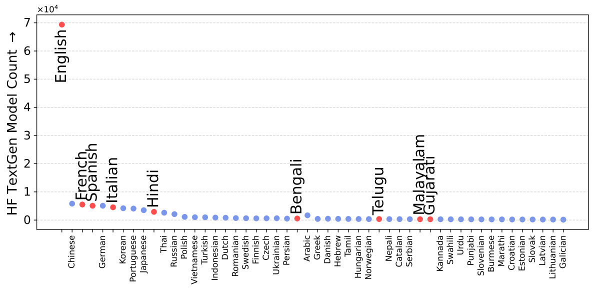


Figure 1: Distribution of publicly available text generation models on HuggingFace (HF) website as of October 22nd, 2025.

roads of the challenges in fluency and factuality. Consequently, we can expect that hallucinations will look different for languages for which models tend to be less fluent (Vazquez et al., 2025) or domains that require more specialized knowledge (George and Stuhlmueller, 2023). Indicators of disparities in access to NLP technology can be illustrated by the variety of tools available for a given language. As we show in Figure 1, the vast majority of languages are undeserved causing a clear imbalance between e.g. French and Telugu or Bengali.

In this paper, we introduce **CAP**, a multilingual question-answering dataset designed to evaluate hallucinations along both fluency and factuality dimensions across nine languages, using scientific publications retrieved from the ACL Anthology. CAP comprises a total of 100 unique questions per language, with eight LLM-generated response annotations per question (Table 1). We conduct extensive experiments to evaluate the performance of six representative hallucination detection baselines spanning reference-based and reference-free paradigms. Our results indicate that existing hallucination detection tools are generally ill-suited for CAP, with most configurations performing near or below random. In addition to benchmarking model performance, we leverage CAP's fine-grained annotation layers to investigate the root causes of hallucinations: in particular, we review the effects of citation counts (as a proxy for notoriety), question type, and context complexity. Our analyses show that linguistic cues such as the type of questions asked or the context used as input for hallucination detection tools tend to explain hallucination rates better than citation counts.

Overall, the contributions of this paper include:

- **CAP (Confabulations from ACL Publications)**, a novel scientific hallucination dataset comprising question-answer pairs for nine languages, including five high-resource and four low-resource languages; and
- **A benchmark evaluation of existing hallucination detection tools** with respect to our proposed definition of hallucination phenomena, showcasing the remaining challenges and suggesting directions for future research in hallucination detection.

## 2. Related Works

LLMs often generate hallucinations: outputs that appear coherent and well-formed but contain factual inaccuracies. As discussed in several recent surveys (Ji et al., 2023; Huang et al., 2024), this issue raises concerns about the reliability of LLMs in knowledge-intensive domains such as scientific writing. In particular, George and Stuhlmueller (2023) highlights how LLMs can generate unsupported claims, invented references, and factually incorrect statements when summarizing or rewriting scholarly articles.

Recent work has focused on factuality evaluation, aiming to assess how well model outputs align with reference information. Early benchmarks (Kryściński et al., 2019; Wang et al., 2020) assessed factual consistency in English summarization using entailment or QA-based proxies. Subsequent studies (Wadden et al., 2022a,b) showed that domain-specific factuality, particularly in scientific writing, cannot be reduced to textual similarity but requires grounding in specialized terminology and evidence retrieval. Qi et al. (2023) analyzed the cross-lingual consistency of factual knowledge in multilingual LLMs, finding that many models achieve low factual alignment across languages and rely heavily on lexical overlap rather than language-independent representations. Together, these findings reveal that current LLMs struggle to maintain factual coherence across scientific domains and languages.

Several benchmarks have been proposed to evaluate hallucination in LLMs, such as Mickus et al. (2024) for general NLG tasks and Yasunaga et al. (2019); Wadden et al. (2022a) for scientific summarization and claim verification. However, all these resources are limited to English. Vazquez et al. (2025) denotes a recent effort toward evaluating hallucination in multilingual environments, reflecting the increasing attention to extend factuality assessment to less represented languages. Building on Vazquez et al.'s work, Rykov et al. (2025) propose a large-scale and multilingual dataset of automatically annotated annotations and demonstrate that such a resource can bolster the efficacy of hallucination detection tools.

## 3. CAP Dataset

### 3.1. Overview

The CAP dataset is a multilingual resource designed to evaluate hallucinations in LLM outputs for scientific text generation. It comprises nine languages in total: five high-resource languages, including English (en), French (fr), Hindi (hi), Italian (it), and Spanish (es); and four low-resource Indic languages, including Bengali (bn), Gujarati (gu), Malayalam (ml) and Telugu (te).


"index": 600,

"title": "RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization",

"abstract": "This study addresses the challenge of extending Large Language Models (LLMs) to non-English languages..",

"doi": "10.18653/v1/2024.acl-long.833",
"url": "https://aclanthology.org/2024.acl-long.833.pdf",
"extracted": true,
"datafile": "2024.acl.xml",
"authors": [{"first": "Jaavid", "last": "J"}, {"first": "Raj", "last": "Dabre"},...],

"question": "ഈ പരീക്ഷയിൽ ഉപയോഗിക്കുന്ന ഇന്ത്യൻ ഭാഷകൾ ഏതൊക്കെയാണ്?"
"model_id": "VishnuPJ/MalayaLLM_7B_Instruct_v0.2",

"model_config": "k50_p0.95_t0.2",
"lang": "malayalam",

"prompt": "J,Jaavid et al. എഴുതിയ 'RomanSetu: Efficiently unlocking multilingual capabilities of Large Language Models via Romanization' എന്ന ലേഖനത്തിൽ.ഈ പരീക്ഷയിൽ ഉപയോഗിക്കുന്ന ഇന്ത്യൻ ഭാഷകൾ ഏതൊക്കെയാണ്?",

"output_text": "ഉത്തരം:1. ഗുജറാത്തി2. തെലുങ്ക്3 ...",

"output_tokens": ["<0x0A>", "<0x0A>", "ഉ", "ത്തരം", ":", "<0x0A>", ...],
"output_logits": [14.4688854218, 16.9766082764,...],

"has_fluency_mistakes": "y",
"has_factual_mistakes": "y"


Figure 2: Example instance extracted from CAP Malayalam. The question roughly translates to *"Which Indian languages are used in this test set?"*. It provides all the metadata regarding the creation of this example including information related to the question and the response generated.

Figure 2 presents an example instance from CAP Malayalam data split. Each instance in the dataset corresponds to a question-answer pair associated with a scientific natural language processing (NLP) paper. These are provided with a comprehensive set of metadata fields, when available, including an index identifier, paper title, abstract, DOI, and URL, as well as all the authors' information, including given name and surname recorded separately to facilitate prompt construction. In addition to the metadata about the question, we also provide metadata about the generated answer (referred to as `output_text` in Figure 2) from the LLM and the LLM model. This includes a tokenized version of the generated answer, the associated logits, and, for the LLM model, we provide the model identifier, model configuration, and the prompt used. Each instance contains two labels, namely, `has_fluency_mistakes` (corresponding to fluency) and `has_factual_mistakes` (corresponding to hallucination) for the generated answer. Altogether, this information enables detailed investigations of model behavior across languages and scientific domains.

### 3.2. CAP Creation Workflow

The data creation process comprises three main stages. First, we curate a cohort of scientific papers. Next, human annotators manually compose questions based on these papers, which are subsequently presented to LLMs to generate corresponding answers.

**Scientific Paper Collection:** We began by *collecting papers* from the ACL Anthology, compiling a set of 293 award-winning papers in NLP from 1995 to 2024. These papers are more likely to be cited and discussed, making them a strong source for hallucination detection. For each selected paper, we extracted available metadata, including the title, abstract, DOI, URL, and list of authors. This structured metadata provides a comprehensive representation of each paper and serves as the foundation for subsequent question creation and answer generation steps.

**Question Creation:** For each language, annotators were provided with a randomly selected subset of $100$ papers (in English) and were instructed to manually create one question for each paper in their assigned language. Since the sampling was performed independently for each language, the resulting subsets differ, meaning that the list of papers annotated in one language does not necessarily overlap with those used in another. For any paper appearing in multiple languages, questions are not simple translations; rather, they are independently crafted by annotators in each language to reflect language-specific perspectives and nuances. Annotators were supported by a script to record each question, providing the corresponding paper link.

**Response Generation using LLMs:** For each question, we generated multiple responses (ranging between 6 to 18) using publicly available, instruction-tuned LLMs capable of handling the target languages (see details in Section 3.4). Multiple outputs per question were obtained by varying generation hyperparameters, including top-$p$, top-$k$, and temperature. Specifically, we use: (1) default setting; (2) top-$p$=0.9, top-$k$=50, temperature=0.1; (3) top-$p$=0.95, top-$k$=50, temperature=0.2. We also vary the input context provided in the prompt used for answer generation by optionally adding the abstract as an additional information.

### 3.3. Annotation Details

**Human Annotators:** Question creation and LLM-generated answer annotations were carried out by a team of ten annotators. All of them are native speakers of the assigned language and possess strong NLP backgrounds, ranging from graduate students to postdoctoral researcher, ensuring domain expertise in the annotation process. The number of annotators across languages is as follows: English (3), French (2), Hindi (2), Italian (1), Spanish (2), Bengali (1), Gujarati (1), Malayalam (1), and Telugu (1). Annotators were instructed to formulate a question using any combination of the title, abstract, full paper, or author information, based on their preference.

> In the article titled [`title`] by [`last`], [`first`][`aux`], [`question`]
>
> Here is the start of the article abstract for your reference: [`abstract`]

Figure 3: Standardized template used for response generation in English, analogous templates are used for other languages.

**Prompt Design:** This process involved carefully designing the question field for each record, which contains a natural-language query in the context of the NLP paper's scientific content. Each question is paired with a prompt, built using a standardized template (see Figure 3).This structure ensures that every query remains clearly anchored to the cited source, providing consistent contextual framing across languages and papers.

**Annotation Procedure:** Due to the large amount of generated answers, we sampled a subset of generated answers for *manual annotation*. The sampling procedure ensures that each output is annotated only once while maintaining a balanced distribution across questions, prompts, and models. This is accomplished by randomly selecting instances and setting target annotation counts for each question–prompt–model combination, aiming for a total of eight annotations per question.

Each selected output was annotated by the same expert who created the question and evaluated along two dimensions: *factual correctness* and *fluency*. Factual correctness was labeled as *yes* or *no*, indicating whether the answer contained factual errors. Fluency was assessed on a three-point scale: *yes* (well-formed), *minor* (minor language issues), or *no* (ungrammatical or disfluent).

During the post annotation sanity check step, several samples containing duplicates were removed from en, es, hi, bn, and gu. This step resulted in final dataset distribution corresponding to Table 1 still keeping unique number of question as 100 for all the nine languages.

### 3.4. CAP Statistics

Table 1 presents the statistics of the CAP dataset across the different splits and languages. In the table, `Q` denotes the number of unique questions created by annotators per language and `R` denotes the number of LLM generated response annotations after post-processing step. For each language, we collect 100 questions and over 500-1000 annotations of generated answers per language. Additionally, we provide train-val-test split to support future research and ensure compatibility with the traditional shared task framework.

| **lang** | **TOTAL** | | **TRAIN** | | **VAL** | | **TEST** | |
|---|---|---|---|---|---|---|---|---|
| | Q | R | Q | R | Q | R | Q | R |
| en | 100 | 588 | 40 | 108 | 30 | 240 | 30 | 240 |
| es | 100 | 660 | 40 | 180 | 30 | 240 | 30 | 240 |
| fr | 100 | 1000 | 40 | 520 | 30 | 240 | 30 | 240 |
| hi | 100 | 905 | 40 | 425 | 30 | 240 | 30 | 240 |
| it | 100 | 1000 | 40 | 520 | 30 | 240 | 30 | 240 |
| **ml | 100 | 788 | — | — | — | — | 100 | 788 |
| **te | 100 | 800 | — | — | — | — | 100 | 800 |
| **bn | 100 | 798 | — | — | — | — | 100 | 798 |
| **gu | 100 | 800 | — | — | — | — | 100 | 800 |

Table 1: Dataset statistics per language. The data are presented using a train/validation/test split following the setup of the shared task introduced in Sinha et al. (2025), for which the data was created.

Table 2 presents all the different LLMs utilized in generating the response for the 100 questions created per language. We selected LLMs with parameter sizes ranging from 2B to 13B. All languages except French and Italian used two LLMs, whereas these used three different models.

Figure 4 illustrates the joint distribution of factuality and fluency annotations against each other for all the languages covered in the CAP data set.

The formulation we adopt here is that *a hallucination is defined as a response that is fluent but false* — that is, linguistically coherent yet factually inconsistent or unsupported with respect to the underlying publication (Bhamidipati et al., 2024). This perspective reframes hallucination detection as a dual-facet problem, requiring models to jointly assess both the fluency and factuality of generated

| Lang. | HF identifier | N. train | N. val. | N. test |
|---|---|---|---|---|
| en | lmsys/vicuna-7b-v1.5 | 42 | 121 | 120 |
| | meta-llama/Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) | 66 | 119 | 120 |
| es | Iker/Llama-3-Instruct-Neurona-8b-v2 | 100 | 120 | 120 |
| | meta-llama/Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024) | 80 | 120 | 120 |
| fr | bofenghuang/vigogne-2-13b-chat | 174 | 78 | 81 |
| | occiglot/occiglot-7b-eu5-instruct | 169 | 75 | 84 |
| | meta-llama/Meta-Llama-3.1-8B-Instruct | 177 | 87 | 75 |
| hi | Cognitive-Lab/LLama3-Gaja-Hindi-8B-v0.1 | 225 | 120 | 120 |
| | sarvamai/OpenHathi-7B-Hi-v0.1-Base | 200 | 120 | 120 |
| it | sapienzanlp/modello-italia-9b | 172 | 86 | 81 |
| | google/gemma-2-9b-it | 183 | 81 | 79 |
| | meta-llama/Meta-Llama-3.1-8B-Instruct | 165 | 73 | 80 |
| **ml | VishnuPJ/MalayaLLM-7B-Instruct-v0.2 | — | — | 395 |
| | sarvamai/sarvam-1 | — | — | 393 |
| **te | meta-llama/Llama-3.1-8B-Instruct | — | — | 399 |
| | Telugu-LLM-Labs/Indic-gemma-7b-finetuned-sft-Navarasa-2.0 | — | — | 399 |
| **bn | BanglaLLM/BanglaLLama-3-8b-bangla-alpaca-orca-instruct-v0.0.1 | — | — | 420 |
| | BanglaLLM/Bangla-s1k-qwen-2.5-3B-Instruct | — | — | 378 |
| **gu | GenVRadmin/AryaBhatta-GemmaUltra-Merged | — | — | 600 |
| | GenVRadmin/AryaBhatta-GemmaGenZ-Vikas-Merged | — | — | 200 |

Table 2: LLMs used for generating responses (R) for each language. ** languages contain only test subset.

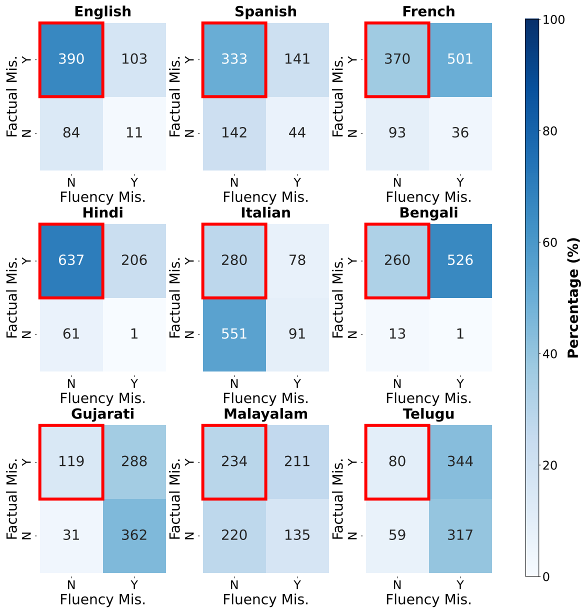


Figure 4: Label distribution per language. Mis. denotes *Mistakes*.

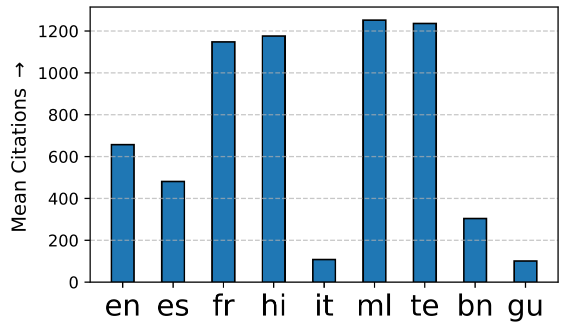


Figure 5: Mean citation per language.

content.

A fact that immediately arises from observing Figure 4 is that *challenges vary from language to language*: for English, Spanish, French, Hindi and Bengali, models struggle to output factually correct responses; for Gujarati and Telugu, the issue is first and foremost fluency; Italian displays mostly responses that are fluent and factual, whereas Malayalam outputs are more uniformly distributed across all four possible cases. Hallucinations proper—i.e., outputs that are fluent but not factual—therefore have likely distinct root causes across languages.

## 4. Characteristics of CAP Dataset

CAP presents a unique and challenging dataset. To examine the factors contributing to its difficulty, we analyze citation counts (Section 4.1) and assess how question types influence the factuality of model outputs (Section 4.2).

### 4.1. Effects of Citation Counts

An interesting side-effect of focusing on scientific publication is that we can provide rough estimates of the popularity of a certain topic, by means of bibliometric indicators. Figure 5 depicts the distribution of mean total citation for each language split from the CAP data set. For example, Malayalam language data contain questions associated to NLP papers that had overall the most citations in comparison to all other languages. On the other hand, Italian and Gujarati contains questions that were created from papers which had relatively the least mean citations.

Interestingly, citation counts do not seem to be indicative of hallucination. Figure 6 shows the distribution of log1plus–transformed citation counts per language, according to whether the output is deemed a hallucination or not. As is apparent, distribution of citation counts are extremely similar for hallucinations and non-hallucinations alike. In fact,

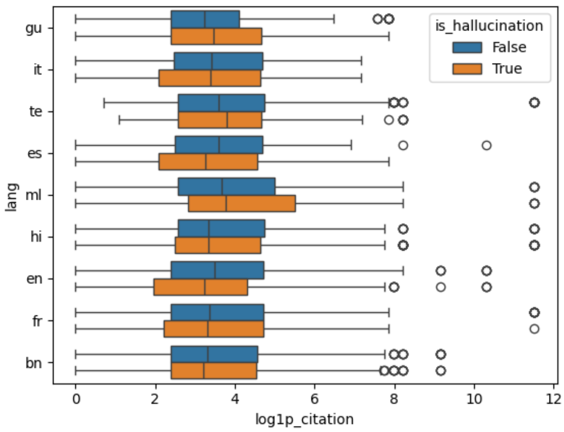


Figure 6: Distribution of citation counts per language, hallucination vs. non-hallucinations.

language-specific Mann-Whitney U-tests comparing the citation counts in hallucinations vs. non-hallucinations suggest no statistically significant difference after Bonferroni correction.

## 4.2. Effects of Question Types

A major confounding factor in Figure 4 is that the questions asked vary across languages, which certainly influences the outputs. To assess this point, Figure 7 presents a comparison of the distribution of the different types of questions based on Graesser and Person (1994)'s taxonomy. Specifically designed for question-answering and information retrieval tasks, this taxonomy categorizes questions into 18 distinct types, taking into account both the format of the expected answer (short or long) and its illocutionary function, such as *verification*, *definition*, *example*, or *comparison* (see also Pomerantz (2005)). To automatically identify the question type for every question created by the annotators, we employed the LLM-as-a-judge approach (Li et al., 2024), using `google/gemma-3-27b-it` (Team et al., 2025) as the LLM judge.

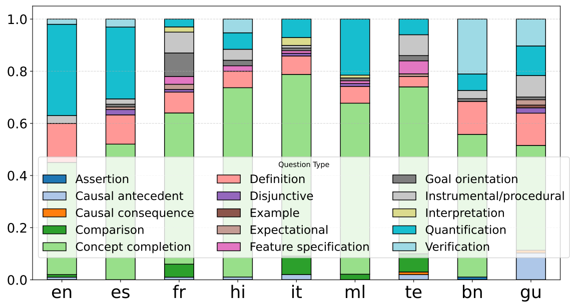


Figure 7: Question Type Distribution per language.

Firstly, we observe that, out of the 18 classes, only 15 classes were identified across the entire dataset, with "*Concept completion*" being the most frequent question type across all nine languages. The corresponding breakdown is displayed in Figure 7, which also highlights the imbalance of question types in the CAP dataset.

| Question Type | # N | # Y | N residual | Y residual |
|---|---|---|---|---|
| Assertion | 0 | 1 | $-0.579$ | $0.411$ |
| Causal antecedent | 9 | 8 | $\mathbf{1.383}$ | $-0.982$ |
| Causal consequence | 0 | 2 | $-0.819$ | $0.581$ |
| Comparison | 11 | 11 | $\mathbf{1.335}$ | $-0.948$ |
| Concept completion | 146 | 360 | $-1.814$ | $1.288$ |
| Definition | 31 | 50 | $0.738$ | $-0.524$ |
| Disjunctive | 3 | 4 | $0.427$ | $-0.303$ |
| Example | 1 | 0 | $\mathbf{1.148}$ | $-0.815$ |
| Expectational | 3 | 3 | $0.697$ | $-0.495$ |
| Feature specification | 2 | 10 | $-1.009$ | $0.716$ |
| Goal orientation | 14 | 4 | $\mathbf{3.243}^{*}$ | $-\mathbf{2.303}^{*}$ |
| Instru./procedural | 17 | 20 | $\mathbf{1.305}$ | $-0.927$ |
| Interpretation | 2 | 4 | $-0.008$ | $0.006$ |
| Quantification | 29 | 92 | $-1.816$ | $1.289$ |
| Verification | 26 | 14 | $\mathbf{3.438}^{*}$ | $-\mathbf{2.442}^{*}$ |

Table 3: Observed counts (N: no hallucination; Y: hallucination) and standardized residuals for each question type. Bold residuals indicate $|Z| \geq 1.96$ ($p < 0.05$). Per question type, $^{*}$ values exhibit statistically significant deviations from expected (non-)hallucination rates.

Next, we also highlight that the type of question asked could have an effect on factuality. To examine this relationship, we perform a Chi-square test between the hallucination label (*yes* or *no*) and the question type, as determined through the LLM-as-a-judge evaluation, shown in Table 3. For each question type and hallucination label, residuals are computed as the difference between observed and expected counts, normalized by the square root of the expected count[1]. The test reveals a significant association between question type and hallucination occurrence, $\chi^2$(14, N=877) = 61.21, $p < .001$. The effect size, measured by Cramér's V, was $0.26$, indicating a small-to-medium effect according to Cohen's guidelines.

In detail, hallucinations are not uniformly distributed across question types: *Verification* and *Goal orientation* questions exhibited strong positive residuals for non-hallucinated responses and negative residuals for hallucinated responses, indicating these types are significantly less prone to hallucinations. Conversely, *Quantification* and *Concept completion* showed moderate positive residuals for hallucination, suggesting a higher hallucination rate in these categories, while most other question types showed weak or no significant deviation from expected frequencies.

[1]For each question type $i$ and hallucination label $j$, the residuals are computed as $r_{i,j} = \frac{O_{i,j}-E_{i,j}}{\sqrt{E_{i,j}}}$ where $O_{i,j}$ is the observed count and $E_{i,j}$ is the expected count which is calculated as $\frac{(\text{total questions of type } i)\cdot(\text{total questions with hallucination label } j)}{N}$.

Together, these findings underscore the richness of the CAP dataset. By reflecting variations in citation-based content and question-type sensitivity, CAP presents a more challenging and realistic benchmark for assessing factual consistency in scientific question answering.

## 5. Benchmarking on CAP

Following Section 4, that highlighted the complexity of CAP, this section evaluates the performance of existing hallucination detection tools on CAP.

### 5.1. Task Formulation

We frame hallucination detection as a binary classification task: given a pair consisting of a scientific publication segment (premise) and an LLM-generated answer (hypothesis), the model must determine whether the hypothesis constitutes a scientific hallucination.

### 5.2. Evaluation Metrics

To ensure comparability across models and languages, we adopted a standardized processing pipeline. Each question-answer pair was segmented into smaller passages to fit model context windows, and baseline models processed each passage independently. The highest probability across passages was used as the final prediction label. We report Macro-F1 scores per language, along with the overall average for all languages.

### 5.3. Models

We benchmark on six representative baselines spanning both *reference-based* and *reference-free* hallucination detection paradigms.

**HHEM-2.1-Open** The Hallucination and Hallucination Evaluation Model (HHEM-2.1-Open; hereafter referred to as HHEM) (Bao et al., 2024) is an open-source model trained on factual consistency datasets to detect hallucinations in long-form outputs generated by LLMs. It functions as a strong factuality classifier, optimized using contrastive, entailment-based training objectives.

**mDeBERTa-v3-base-xnli** (Laurer et al., 2022) is a multilingual entailment model fine-tuned on 2.7 million natural language inference (NLI) examples spanning over 100 languages. Aligned with our task formulation, the model evaluates whether a generated answer *entails*, *contradicts*, or is *neutral* with respect to the source publication. Outputs labeled as *contradiction* or *neutral* are considered hallucinations.

**SelfCheckGPT** (Manakul et al., 2023) is a reference-free hallucination detector that operates without access to external context. It estimates factual consistency by sampling multiple outputs from a language model and measuring the degree of internal agreement among them. We utilized a `google/gemma-2-9b` model for this, and therefore denote is as SelfCheckGemma.

**XLM-RoBERTa-XL Hallucination Detector** This multilingual reference-based hallucination detector (Bondarenko, 2024) is built on the XLM-RoBERTa-XL backbone. The model casts hallucination detection as a binary classification task, employing a self-adaptive hierarchical encoder fine-tuned in two stages: contrastive learning to optimize sentence embeddings, followed by supervised fine-tuning for classification.

**FAVA** (Mishra et al., 2024) is a reference-based model for fine-grained hallucination detection and correction. It employs an editor language model trained on synthetically generated data to identify and revise factual inaccuracies in generated text.

**HDM-2** (Paudel et al., 2025) is a comprehensive hallucination detection model designed to validate outputs from LLMs using both contextual evidence and common knowledge. It employs a multi-task architecture comprising separate modules for context-based and common-knowledge verification, and generates hallucination scores at both the sentence and token levels.

All baselines utilize pretrained large language models in zero-shot inference setting without additional fine-tuning for fair comparison.

### 5.4. Results

**Disentangling the effects of fluency** We first seek to establish the performances of existing models. In Table 4a, we summarize macro-F1 scores for the 6 baselines described above. A related point of interest is the impact of the definition of hallucination we adopt in this work, as outputs that are fluent, yet not factually correct. To disentangle the effects of fluency, in Table 4b we also report performances on the subset of datapoints annotated as fluent, i.e., simplifying the task to one of factual correctness classification.

A few observations can be drawn from results in Table 4. First, off-the-shelf models tend to perform remarkably poorly. Overall performances across all test datapoints are always below $0.46$. Given that we frame the problem as a binary classification, this result puts into question the efficacy of hallucination detectors in out-of-domain settings.

| Model | en | es | fr | hi | it | ml | te | bn | gu | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| HHEM | 0.486 | 0.420 | 0.357 | 0.446 | 0.258 | 0.366 | 0.093 | 0.247 | 0.249 | 0.294 |
| mDeBERTaNLI | 0.435 | 0.497 | 0.586 | 0.463 | 0.535 | 0.316 | 0.214 | 0.500 | 0.254 | 0.389 |
| SelfCheckGemma | 0.503 | 0.494 | 0.520 | 0.556 | 0.394 | 0.454 | 0.409 | 0.494 | 0.328 | 0.459 |
| XLMRobertaXL | 0.440 | 0.417 | 0.243 | 0.405 | 0.204 | 0.390 | 0.136 | 0.407 | 0.278 | 0.334 |
| FAVA | 0.433 | 0.525 | 0.312 | 0.490 | 0.271 | 0.430 | 0.380 | 0.508 | 0.357 | 0.456 |
| HDM2 | 0.459 | 0.425 | 0.277 | 0.443 | 0.276 | 0.395 | 0.298 | 0.524 | 0.260 | 0.398 |

(a) Macro-F1 metric, considering both classes.

| Model | en | es | fr | hi | it | ml | te | bn | gu | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| HHEM | 0.467 | 0.465 | 0.431 | 0.554 | 0.302 | 0.543 | 0.411 | 0.529 | 0.400 | 0.514 |
| mDeBERTaNLI | 0.422 | 0.458 | 0.542 | 0.394 | 0.552 | 0.402 | 0.397 | 0.472 | 0.512 | 0.518 |
| SelfCheckGemma | 0.518 | 0.457 | 0.589 | 0.519 | 0.384 | 0.519 | 0.577 | 0.498 | 0.480 | 0.530 |
| XLMRobertaXL | 0.475 | 0.424 | 0.413 | 0.468 | 0.225 | 0.357 | 0.359 | 0.486 | 0.452 | 0.407 |
| FAVA | 0.453 | 0.538 | 0.468 | 0.671 | 0.305 | 0.411 | 0.509 | 0.461 | 0.465 | 0.475 |
| HDM2 | 0.493 | 0.436 | 0.426 | 0.458 | 0.317 | 0.419 | 0.393 | 0.548 | 0.503 | 0.460 |

(b) Macro-F1 metric, considering only fluent datapoints.

Table 4: Performance comparison between hallucination detectors across languages.

Second, simplifying the problem to that of a classification of factuality among fluent outputs need not yield consistent improvements. While we observe improvements for low resource Indic languages, the same does not hold consistently for high resource languages, and overall performances remain remarkably poor. In short, the low scores we observe when measuring performance across all items are not only due to the lesser regularity of non-fluent datapoints, but rather reflects genuine difficulty intrinsic to the CAP dataset.

| | Model | F1 | Rec | Prec |
|---|---|---|---|---|
| en | FAVA | −0.026 | −0.056 | −0.023 |
| | HDM2 | 0.118 | 0.061 | 0.016 |
| | XLMRobertaXL | 0.019 | 0.009 | 0.502 |
| te | FAVA | 0.157 | 0.059 | 0.035 |
| | HDM2 | 0.115 | 0.034 | 0.001 |
| | XLMRobertaXL | 0.006 | 0.003 | 0.000 |

(a) All datapoints

| | Model | F1 | Rec | Prec |
|---|---|---|---|---|
| en | Fava | −0.047 | −0.019 | −0.005 |
| | HDM2 | 0.161 | 0.171 | 0.050 |
| | XLMRobertaXL | 0.043 | 0.022 | 0.503 |
| te | Fava | −0.096 | −0.014 | 0.035 |
| | HDM2 | 0.088 | 0.020 | 0.068 |
| | XLMRobertaXL | 0.000 | 0.000 | 0.000 |

(b) Only fluent datapoints

Table 5: Effects of narrowing the context (scores when using the relevant context as input minus scores when using the full paper).

**Effects of long-context input** Results in Table 4 are obtained by selecting the highest probability across text chunks[2] of the entire article. While this approach is practically motivated, we can also expect it biases the models towards classifying outputs as hallucinated. This challenge is intrinsic to the long-context nature of the reference documents.

To assess this point more formally, we perform an ablation study: we compare performances on a subset of items when feeding the entire article as opposed to what we would obtain by using as reference only the section deemed relevant by the annotators. To balance annotation efforts and coverage, we focus on one high resource (English) and one low resource language (Telugu). We consider three reference-dependent models: Fava, HDM2 and XLMRobertaXL.

Results are listed in Table 5, presented as the margin of improvement when using as input the relevant section only. Following our previous approach, we report both scores on the full dataset (in Table 5a) as well as scores obtain when considering only items presenting no issues in fluency (in Table 5b). Results suggest that performance improvements are highly contingent on the exact model considered: while Fava systematically yields lower recall and macro-F1 scores, HDM2 generally benefits from more targeted contexts. This suggests that not all models are equally sensitive to irrelevant elements of context.

[2] Chunks are created using whitespace-based segmentation, kept well below the model context window to account for subword tokenization, and include overlaps to preserve contextual continuity.

## 6. Conclusion

We present a novel scientific hallucination dataset, the CAP dataset, comprising question–answer pairs across five high-resource and four low-resource languages. Each instance is annotated for both factuality and fluency, enabling a comprehensive evaluation of LLMs in multilingual scientific contexts. This dataset extends the scope of hallucination research to the scientific domain and promotes cross-lingual analysis of factual consistency. CAP directly addresses the growing tendency of LLMs to produce fluent yet factually incorrect content, offering a challenging benchmark for evaluating factual grounding.

Our focus on evaluating hallucination as a phenomenon at the crossroads of fluency and factuality offers an interesting complementary view to other taxonomies of hallucination. For instance, while the definition we rely on, in and of itself, could apply to extrinsic and intrinsic hallucinations alike (i.e., whether the hallucination directly contradicts the given input text, vs. whether it contradicts the LLM's training data; Ji et al., 2023; Bang et al., 2025), the dataset we provide relies mostly on knowledge obtained through exposure during pretraining or instruction-tuning (as we do not provide the entire contents of papers to the LLMs we assess) — meaning we have focused primarily on extrinsic hallucinations. Given this state of affairs, it is worth highlighting that our extrinsic indicator of topic popularity (viz. citation counts) does not appear to be a factor impacting hallucination rates, whereas intrinsic linguistic cues (such as question types) do impact LLM outputs in a measurable way. On the other hand, the data we collect provides evidence that fluency and factuality play different roles for different languages, whereas our benchmarking experiments underscore the difficulty inherent to processing noisier, less-fluent outputs. Our dual-faceted outlook on hallucination therefore provides an interesting complementary lens to reassess prior research in the field.

Future work may include expanding the dataset to additional scientific disciplines such as the medical domain (Pal et al., 2023) and extremely low resource languages (Joshi et al., 2020), exploring the interaction of hallucination with unseen languages, and leveraging CAP for developing robust, hallucination-aware generation models.

## 7. Acknowledgments

The work described herein has been supported in part by the EC under the grant No. 101195233 (OpenEuroLLM).

## 8. Limitations and Ethical Considerations

Some limitations of our study concern both terminology and data scope. First, the use of the term *hallucination* to describe AI-generated factual errors is inherently metaphorical and may be misleading, as it implies flawed perception rather than the statistical generation processes that underlie large language models (Hicks et al., 2024). Although we adopt this term for consistency with existing literature, we acknowledge its conceptual limitations and the potential influence such framing may have on public understanding of AI reliability. Second, the CAP dataset is limited to NLP papers from ACL venues, which may restrict the generalizability of the findings to other domains. Nevertheless, this work represents an initial step toward opening a new research direction. Thirdly, while our dataset spans nine languages, coverage is uneven, with lower representation and translation quality in low-resource languages. Finally, despite rigorous design and review, human annotations may still reflect subjective judgments, especially in cross-lingual assessments of factual correctness. In addition, because each question was annotated by a single annotator, the annotations may further reflect individual subjectivity and do not allow us to report inter-annotator agreement.

From an ethical perspective, the CAP dataset is derived from publicly available sources and contains no personal data. Nonetheless, model outputs may include scientifically inaccurate or misleading statements; these should not be treated as factual. The dataset is released strictly for research purposes to promote safer and more transparent scientific text generation.

## Bibliographical References

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Patanjali Bhamidipati, Advaith Malladi, Manish Shrivastava, and Radhika Mamidi. 2024. Maha bhaashya at SemEval-2024 task 6: Zero-shot multi-task hallucination detection. In *Proceedings of the 18th International Workshop on Semantic*

*Evaluation (SemEval-2024)*, pages 1685–1689, Mexico City, Mexico. Association for Computational Linguistics.

Ivan Bondarenko. 2024. The reference-based detector of llm hallucinations by ivan bondarenko. https://huggingface.co/bond005/xlm-roberta-xl-hallucination-detector.

Roberts Darģis, Guntis Bārzdiņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

Charlie George and Andreas Stuhlmueller. 2023. Factored verification: Detecting and reducing hallucination in summaries of academic papers. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 107–116, Bali, Indonesia. Association for Computational Linguistics.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew

Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*, 26(2):38.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Bibek Paudel, Alexander Lyzhov, Preetam Joshi, and Puneet Anand. 2025. Hallucinot: Hallucination detection through context and common knowledge verification.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Jeffrey Pomerantz. 2005. A linguistic analysis of question taxonomies. *Journal of the American Society for Information Science and Technology*, 56(7):715–728.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. *arXiv preprint arXiv:2310.10378*.

Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. When models lie, we learn: Multilingual span-level hallucination detection with psiloqa.

Aman Sinha, Federica Gamba, Raúl Vázquez, Timothee Mickus, Ahana Chattopadhyay, Laura Zanella, Binesh Arakkal Remesh, Yash Kankanampati, Aryan Chandramania, and Rohit Agarwal. 2025. SHROOM-CAP: Shared task on hallucinations and related observable overgeneration mistakes in crosslingual analyses of publications. In *Proceedings of the 1st Workshop on Confabulation, Hallucinations and Overgeneration in Multilingual and Practical Settings (CHOMPS 2025)*, pages 70–80, Mumbai, India. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kees van Deemter. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. 2025. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

## Appendix A: Annotation Guidelines

We conceptualize hallucination as a two-dimensional phenomenon, covering factuality and fluency. Formally, we define a hallucination as a *response that is fluent but false* — that is, linguistically coherent yet factually inconsistent or unsupported with respect to the underlying publication.

The annotation task was organized as follows:

1. Question generation: Annotators were presented with the title of a paper (selected from a set of award-winning NLP papers in the ACL Anthology) and were asked to review the paper — considering the title, abstract, and any content they deemed necessary — to formulate a relevant question.

2. Response generation: Annotators then used an LLM of their choice for their assigned language to generate multiple responses. The LLMs were provided with the paper's title, abstract, and the generated question (but not the full paper).

3. Annotation: For each generated response, annotators evaluated the generated response for two dimensions:

    - Factuality: Verifying against the full content of the article corresponding to the question.
    - Fluency: Assessing for orthographic errors, including gibberish, incorrect script, spelling mistakes, or output in the wrong language.

The extended guideline covering all possibilities of hallucination for annotations is provided in Figure 8.

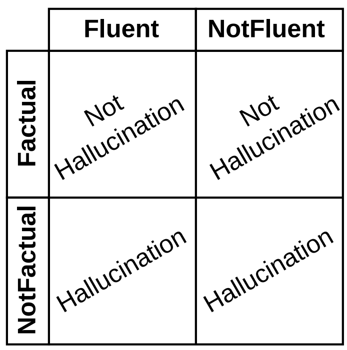


Figure 8: Hallucination Annotation Schema.