PerHalluEval: Persian Hallucination Evaluation Benchmark for Large Language Models

Mohammad Hosseini[1*], **Kimia Hosseini**[1*], **Shayan Bali**[2], **Zahra Zanjani**[1], **Saeedeh Momtazi**[1]
[1]Amirkabir University of Technology, Tehran, Iran
{mohammad, kimia.h, zahra.zanjani99, momtazi}@aut.ac.ir
[2]King's College London, London, UK
shayan.bali@kcl.ac.uk

Abstract

Hallucination is a persistent issue affecting all large language models (LLMs), particularly within low-resource languages such as Persian. **PerHalluEval** (**Per**sian **Hallu**cination **Eval**uation) is the first dynamic hallucination evaluation benchmark tailored for the Persian language. Our benchmark leverages a three-stage LLM-driven pipeline, augmented with human validation, to generate plausible answers and summaries regarding QA and summarization tasks, focusing on detecting extrinsic and intrinsic hallucinations. Moreover, we used the log probabilities of generated tokens to select the most believable hallucinated instances. In addition, we engaged human annotators to highlight Persian-specific contexts in the QA dataset in order to evaluate LLMs' performance on content specifically related to Persian culture. Our evaluation of 12 LLMs, including open- and closed-source models, using **PerHalluEval**, revealed that the models generally struggle to detect hallucinated Persian text. We showed that providing external knowledge, i.e., the original document for the summarization task, could partially mitigate hallucinations. Furthermore, there was no significant difference in terms of hallucination when comparing LLMs specifically trained for Persian with others.

Keywords: Hallucination, Large Language Models, Persian, Benchmark, Evaluation, QA, Summarization

1. Introduction

Large Language Models have rapidly achieved global prominence due to their versatility in Natural Language Processing (NLP) tasks (Naveed et al., 2023), driving extensive usage (Yang et al., 2024b). However, despite their impressive capabilities, a primary challenge affecting all LLMs is their tendency to "hallucinate," contextual misinterpretation, factual fabrication, specificity distortion, incorrect inference, and unwarranted uncertainty (Ji et al., 2023). Consequently, even prominent state-of-the-art models—including GPT-4 (Achiam et al., 2023), and Meta's LLaMA (Touvron et al., 2023)—have all exhibited instances of hallucinations, highlighting that this issue persists even in highly advanced systems (Bang et al., 2025).

On the other hand, although the performance of LLMs on high-resource languages such as English has advanced, more research is needed to thoroughly assess and enhance their performance on low-resource languages, especially those with complex structures and rich morphology. (Chataigner et al., 2024; Zhang et al., 2025) Persian, due to its extensive morphology, pro-drop syntax, Ezafe construction, and right-to-left script, is also regarded as one of these demanding low-resource languages (Shamsfard, 2019; Ghayoomi et al., 2010; Khashabi et al., 2020).

Despite the datasets available for the Persian language (Farsi et al., 2024; Sabouri et al., 2022), the language is not rich in resources. This limitation, along with its grammatical and lexical complexity, makes its study on LLMs, especially on hallucination detection, more challenging. Numerous benchmarks like HalluLens (Bang et al., 2025), ANAH/ANAH-v2 (Ji et al., 2024), GraphEval (Cao et al., 2024), and FactBench (Bayat et al., 2025) are available for assessing hallucinations, yet they predominantly cater to English and other well-resourced languages. The evaluation of hallucinations in Persian has remained largely unaddressed. To date, comprehensive resources for evaluating Persian LLM hallucinations do not virtually exist, underscoring a significant research gap that needs to be addressed to make LLMs more reliable and resilient.

To address this gap, we introduce **PerHalluEval** (**Per**sian **Hallu**cination **Eval**uation), the first dynamic hallucination detection benchmark specifically tailored for Persian. We propose a novel multi-agent pipeline, augmented with human validation, to generate diverse, challenging hallucinated examples by generating two hallucinated datasets based on the PN-Summary (Farahani et al., 2021) and PQuAD (Darvishi et al., 2023) datasets. Moreover, to get more accurate data, we employ a competent LLM in addition to a probabilistic verifier to rigorously filter out low-quality instances. Our approach distinctly categorizes intrinsic hallucinations, i.e., contradictions to the source, from extrinsic hallucinations, i.e., unsupported content, and demonstrates its effectiveness through extensive evalua-

[*]Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3107–3127
11–16 May 2026.

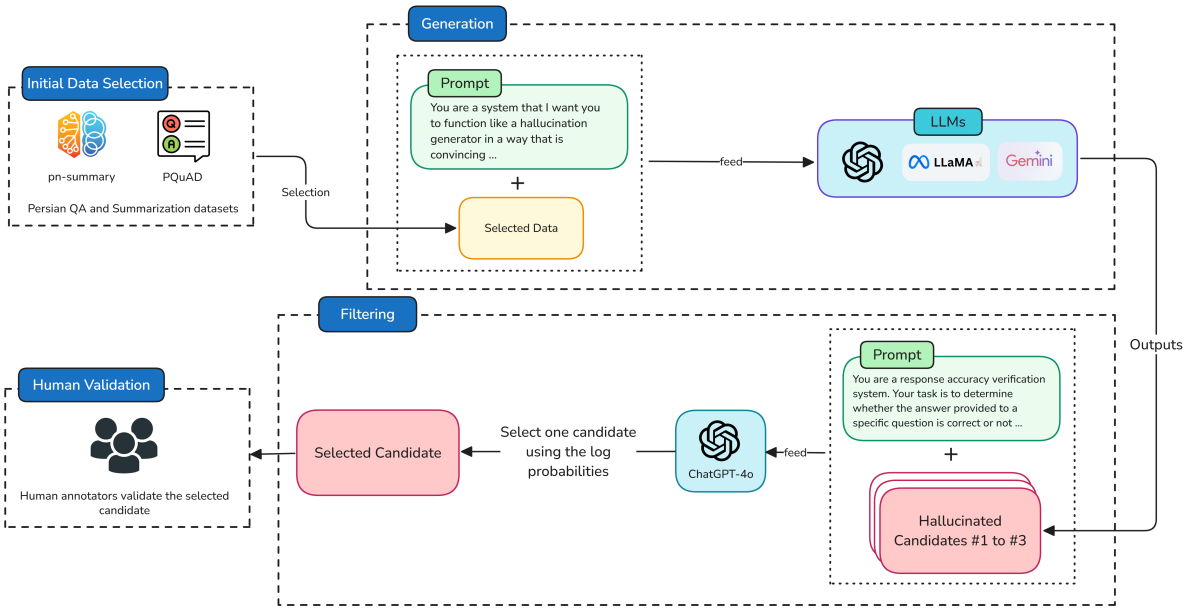


Figure 1: The pipeline of constructing PerHalluEval dataset, consisting of three stages: initial data selection, generation, and filtering, which was augmented with human validation. The visualized prompts are for the QA task, translated into English.

tions of various LLMs, underscoring their unique challenges with Persian linguistic features (Bang et al., 2025).

Following that, we perform a benchmark task to evaluate 12 LLMs on our crafted hallucination dataset—to evaluate their performance in generating reliable outputs in Persian—using our three metrics: Hallucination Recall, Factual Recall, and Hamming Score, which will be described in more detail subsequently. To contextualize these aggregate results, we additionally provide representative qualitative error-analysis cases in Appendix D. Our evaluation of different varieties of models, including models explicitly fine-tuned for Persian and mainstream model families, shows that they struggle to detect hallucinated Persian content.

Our findings aid researchers in identifying hallucinations in models, particularly for the Persian language, and pave the way for future studies as a comprehensive hallucination evaluation benchmark designed for Persian.

2. Related Work

In recent years, hallucinations in LLMs have become a major concern (Huang et al., 2023; Brown et al., 2020). The propensity for hallucination is further exacerbated in specialized domains characterized by high expertise requirements and limited training data, such as circuit design (Petersen et al., 2025; Cassidy et al., 2025; Fayyazi et al., 2024) and network analysis (Hao et al., 2024; Tarzjani and Krishnamachari, 2025). Researchers distinguish between intrinsic hallucinations—when outputs contradict the source—and extrinsic hallucinations, where generated content seems plausible but cannot be verified (Ji et al., 2023; Maynez et al., 2020). While intrinsic errors are often easy to spot, extrinsic ones are much subtler, arising from the model's knowledge gaps or mistaken assumptions (Ji et al., 2023).

To address these challenges, a range of benchmarks has emerged (Huang et al., 2023), including HaluEval (Li et al., 2023), HalluQA (Cheng et al., 2023), ANAH (Ji et al., 2024), HalluDial (Luo et al., 2024), and others (Dziri et al., 2021). These resources reveal weaknesses of LLMs in fact-checking, dialogue, and QA. Newer tools, such as HalluLens, refine the distinction between factuality and hallucination and enable more effective evaluation (Bang et al., 2025). Recent benchmarks like RAGTruth and FactCHD focus on retrieval-augmented and complex reasoning settings (Chen et al., 2023; Kryscinski et al., 2019).

Mitigation strategies now include methods like SelfCheckGPT and contrastive learning (Manakul et al.). QA-based evaluation protocols (QAGS (Wang et al., 2020), FEQA (Durmus et al., 2020)) and correction models (Span-Fact (Dong et al., 2020), FASum (Zhu et al., 2021)) further improve factual consistency.

Despite these advances, most progress has centered on high-resource languages. In contrast, Persian NLP lags behind (Mehrban and Ahadian, 2023; Sadjadi et al., 2024; Abbasi et al., 2023; Rostami et al., 2024). Commonly Persian benchmark resources, such as ParsiNLU (Khashabi et al.,

2020), Persian in a court (Farsi et al., 2025b), Khayyam Challenge (Ghahroodi et al., 2024) and Melac (Farsi et al., 2025a), do not address hallucination (Jolfaei and Mohebi, 2025).

Approaches that work for English or Chinese typically require extensive external datasets, which are unavailable for Persian. To bridge this gap, we introduce PerHalluEval, a fresh, dynamic benchmark designed specifically for Persian. PerHalluEval goes beyond static testing, enabling more effective hallucination detection in both question answering and summarization tasks. By drawing on multiple LLMs and considering the distinct features of the Persian language and culture, our benchmark paves the way for a more reliable evaluation and helps set the stage for future progress in low-resource NLP.

3. PerHalluEval Benchmark

The main goal of constructing this benchmark is to evaluate LLMs' hallucinations in Persian regarding two categories, extrinsic and intrinsic hallucination (Ji et al., 2023). Original correct sentences and their hallucinated ones, which are generated by an LLM-driven pipeline, augmented with human validation, constitute the PerHalluEval benchmark. The construction pipeline, as shown in Figure 1, comprises three stages: initial data selection, generation, and filtering. Human annotators then validate the selected candidate. To avoid saturation through leakage, this LLM-driven benchmark is thought of as dynamic, with data that can be dynamically generated on demand via a generation-and-filtering pipeline, maintaining a continually refreshed pool that resists memorization and preserves evaluation integrity. (Bang et al., 2025).

3.1. Initial Data Selection

The first step in constructing the PerHalluEval dataset is collecting initial samples. For this purpose, two existing Persian datasets: PN-Summary (Farahani et al., 2021) and PQuAD (Darvishi et al., 2023), are selected, covering summarization and question answering tasks, respectively. The Pn-summary dataset contains 93,207 records consisting of articles and their corresponding summaries. The PQuAD dataset includes about 80,000 questions, their corresponding answers, and a related context passage.

In this paper, 4,000 instances are sampled from each dataset using task-specific procedures to preserve diversity and coverage. To select 4,000 questions for the QA task (PQuAD), which has 19 topical categories labeled, stratified sampling is used to ensure that the label distribution matches the original dataset within ±0.3 percentage points. Clustering by document length is also used to ensure coverage from short to long articles for the summarization task (PN-Summary), which does not have topic tags. Each article is encoded using the logarithm of its token count, and then k-means (k = 40, cosine distance) is applied. Articles are then sampled uniformly from each length cluster.

3.2. Generation

In the second stage, the hallucinated version of the data acquired from the previous stage is produced using some well-known LLMs with appropriate instructions. For this purpose, two closed-source LLMs–GPT-4o and Gemini 2.0 Flash—due to their outstanding performance in Persian, along with one open-source LLM, Llama-3.3-70B.

One of the crucial aspects of this pipeline is developing a strong instruction for the mentioned LLMs to generate hallucinated responses. This instruction consists of four parts: an overview of the goal, hallucination patterns, a few-shot example for each pattern, and output structures. Figure 2 demonstrates the instruction structure for the QA task.

The first part of the instruction provides a detailed description of the definition of the task, inputs, outputs, and expected response. Hallucination patterns are scenarios that explain how LLMs must produce hallucinated answers. The same set of diverse patterns is used for both QA and summarization tasks. Five types of patterns are considered for both tasks: contextual misinterpretation, factual fabrication, specificity distortion, incorrect inference, and unwarranted uncertainty (Ji et al., 2023). An example per pattern illustrates a pair of correct and corresponding hallucinated versions. The last part of the instruction includes constraints on the output structure, such as response length.

The prepared instructions are fed into the three mentioned LLMs, accompanied by the curated, accurate data during the initial data selection phase. Finally, in this stage, there are three hallucinated candidates for each of the received samples.

3.3. Filtering Hallucinated Candidates

The objective of the third stage is to obtain the most believable and challenging hallucinated content among the three candidates. Accordingly, a simple prompt format is used, reflecting how most non-expert users normally engage with language models, favoring straightforward prompts over complex engineering techniques (Mishra and Nouri, 2023). Appendix E illustrates the prompt formats for both QA and summarization tasks, accompanied by example inputs.

When this prompt and each candidate are fed into the verifier GPT-4o, it returns "Y" or "N," along with their corresponding log probabilities. Log probabilities are utilized because, as shown in (Kauf

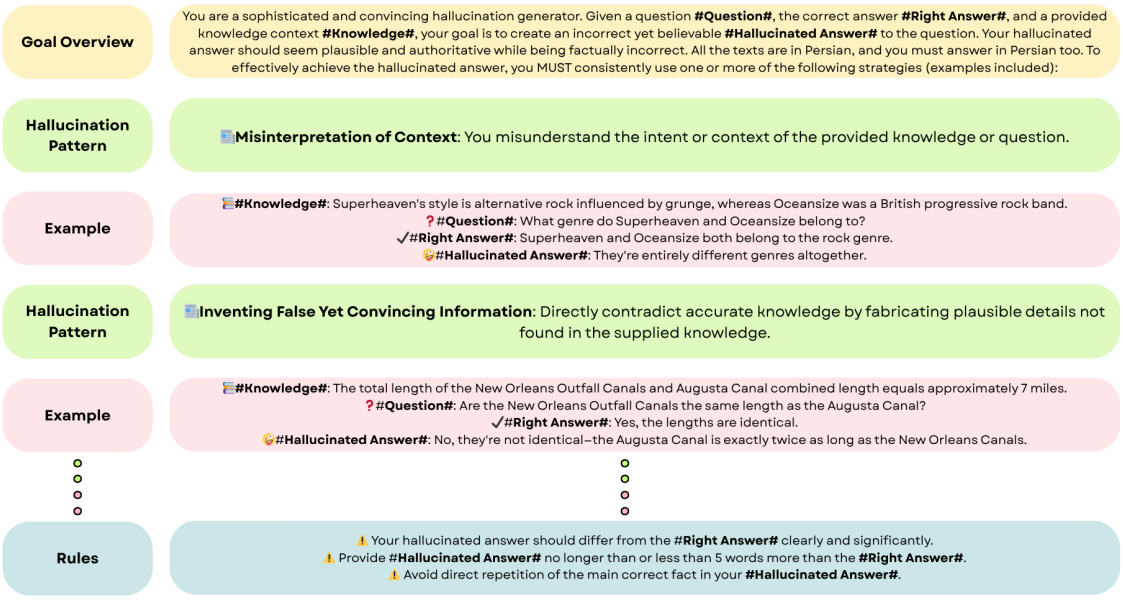


Figure 2: Instruction structure of generating hallucinated content, including an overview of the goal, hallucination patterns, a few-shot example for each pattern, and output structures, for the QA task. The Figure shows the English translation of the original Persian instructions.

et al., 2024), they provide a more reliable assessment of semantic plausibility than direct zero-shot prompting, which frequently produces inconsistent and inferior outcomes.

One of the challenges is that the models' response log probabilities contain different characters corresponding to our 'Y' and 'N' labels. Accordingly, log probabilities are extracted for all 'Y' and 'N' labels and their equivalents using regular expression pattern (as illustrated in Appendix G) to ensure the accuracy of our evaluation. If an output contains no token that matches either of these regexes, the instance is treated as unparseable and excluded from the analysis. To enhance interpretability, a percentage is derived from the log probability by applying the exponential function. To evaluate the candidates' credibility, a confidence margin is calculated using the following score function:

$$Score = P('Y') - P('N') \quad (1)$$

A higher score indicates that the model regards this response as more plausible, with the rise in P('Y') and decline in P('N'). Ultimately, the hallucinated candidate with the greatest score, indicating the most reasonable one to the model, is selected as the final answer, which is the most challenging hallucinated content to be recognized by LLMs. This robust selection is then utilized to evaluate the capability of various LLMs in detecting hallucinatory contents in QA and summarization tasks.

Additionally, a post-hoc analysis of the filtering outputs shows that the finally selected hallucinated candidates are well-distributed across source models (Table 1), suggesting that our selection procedure does not favor any one generator and thus mitigates generator-specific selection bias.

Table 1: Source-model share of selected hallucinations.

Task	Model Name	Share (%)
QA	Gemini 2.0 Flash	40%
	GPT-4o	31%
	Llama 3.3 70B	29%
TS	Gemini 2.0 Flash	38%
	GPT-4o	30%
	Llama 3.3 70B	32%

3.4. Human Validation

Besides the above work on the quality and reliability of the hallucination datasets, we also conducted full human annotation. All 4,000 items in the summarization and QA datasets were independently annotated by three annotators. Annotators followed a concise instruction sheet: label an item as *hallucinated* when the generated content is unsupported by, or contradicts, the source; label it as *factual* when all claims are supported by the source. We used majority voting: an item entered the final hallucination set if at least two annotators labeled it as hallucinated; otherwise, it was excluded. Instructions provided in Appendix C.1.

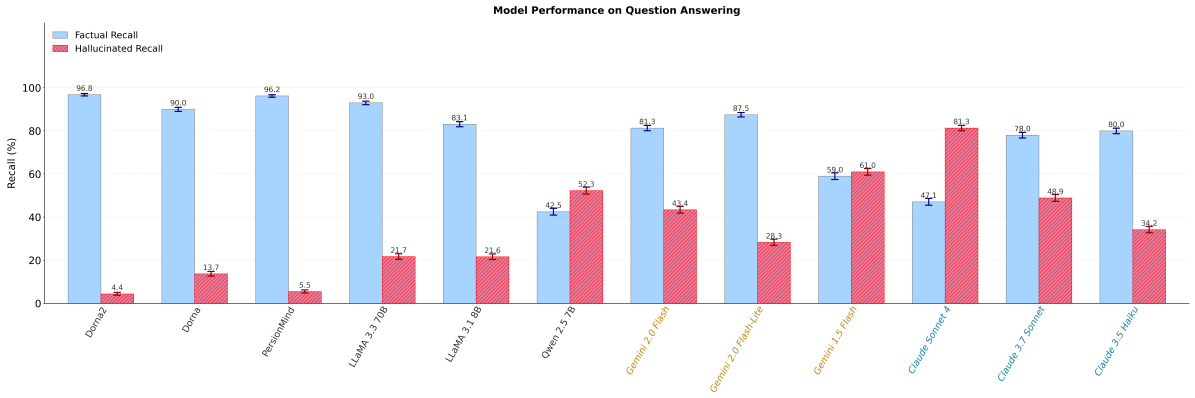


Figure 3: Comparison of evaluated LLMs' performances on the QA task. Error bars show 95% confidence intervals across evaluation samples. Open-source models are displayed in black, Gemini in orange, and Anthropic in blue.

This majority-vote decision rule achieved high retention rates in the two datasets. In detail, 3,829 out of 4,000 QA items and 3,917 out of 4,000 summarization items were retained during validation. The inter-annotator agreement was assessed in terms of Gwet's AC1—reliability index, a measure robust against prevalence and marginal probability bias, and which attained high indices of 0.89 and 0.91 for the QA and summarization set, respectively. These results point to the stable annotation agreement and testify to the validity of the human-curated hallucination labels of the benchmark

3.5. Evaluation

Models. In this benchmark, 12 Large Language Models, spanning a wide range of open-source and commercial models, are evaluated. To address Persian-specific performance, some models that are explicitly fine-tuned on Persian datasets are included, specifically Dorna[1], Dorna2[2], and PersianMind (Rostami et al., 2024). Other multilingual open-source models are Llama 3.1 8B Instruct, Llama 3.1 70B Instruct (Grattafiori et al., 2024), and Qwen 2.5 7B (Yang et al., 2024a). Furthermore, Anthropic's Claude Sonnet 4, Claude Sonnet 3.7, and Claude Haiku 3.5, as well as Google's Gemini family, including Gemini 1.5 Flash, Gemini 2.0 Flash, and Gemini 2.0 Flash-Lite, constitute the evaluated commercial models. For reproducibility, all decoding settings, API/model versions, hardware specifications, and repository links for every evaluated model are documented in Appendix A.

Proposed Evaluation Metrics. To effectively measure model performance in distinguishing between factual (correct) and hallucinated (incorrect) outputs, three targeted measures are introduced for our evaluation: *Factual Recall*, *Hallucinated Recall*, and *Hamming Score*. More info about the Hamming Score can be found in Appendix B.

In each evaluation scenario, each prediction made by a model can be categorized into one of four classical outcomes:

True Positive (TP): the model correctly identifies a factual output (when the given answer/summary is right, the model also predicts 'Y').

False Positive (FP): the model incorrectly predicts hallucinated output as factual (although the given answer/summary is hallucinated, the model predicts 'Y').

True Negative (TN): the model correctly identifies hallucinated output (when the given answer/summary is hallucinated, the model also predicts 'N').

False Negative (FN): the model incorrectly predicts factual output as hallucinated (although the given answer/summary is right, the model predicts 'N'). Based on the above definitions, the metrics are explicitly defined in (2) and (3).

$$\text{Factual Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Hallucinated Recall} = \frac{TN}{TN + FP} \tag{3}$$

The metric *Factual Recall* evaluates the model's reliability in correctly accepting factual information by quantifying the proportion of correctly detected factual outputs among all instances labeled as factual by the dataset. Conversely, the metric *Hallucinated Recall* assesses the capability to precisely detect hallucinated content by measuring the proportion of correctly detected hallucinated outputs within all instances labeled as hallucination in the PerHalluEval.

These two metrics provide complementary evaluations of a model's performance. A high Factual Recall demonstrates the model's effectiveness in

[1]https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct

[2]https://huggingface.co/PartAI/Dorna2-Llama3.1-8B-Instruct

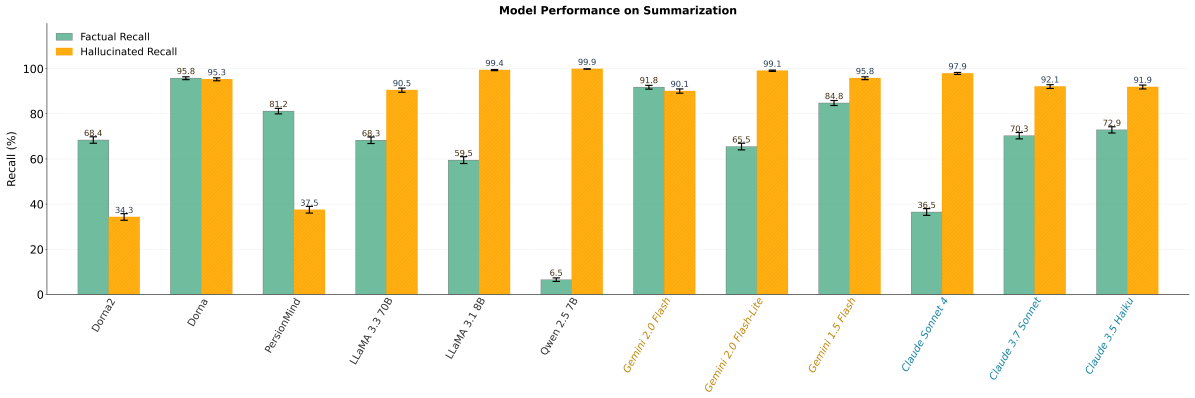


Figure 4: Comparison of evaluated LLMs' performances on the summarization task. Error bars show 95% confidence intervals across evaluation samples. Open-source models are displayed in black, Gemini in orange, and Anthropic in blue.

accepting factual content without mistakenly labeling it as hallucinated data, whereas a high hallucinated Recall demonstrates its robustness in identifying hallucinated responses without labeling them as factual data.

Evaluation Method. To evaluate the selected LLMs on the benchmark, a straightforward prompting approach is adopted that is consistent with the filtering stage described previously. Specifically, the prompts used for evaluating the models on both QA and summarization tasks are identical to those employed in the filtering step. Each prompt is delivered to the models, starting with defining a role for the LLM with the format *"You are a ..."*. Furthermore, each model is separately provided with both the hallucinated and the correct (non-hallucinated) versions of answers/summaries. The models are instructed to produce a binary classification output ('Y' or 'N') for each provided instance. This evaluation setup enables a systematic and consistent comparison of the models' capabilities in distinguishing hallucinated from accurate content.

4. Main Results

4.1. Question Answering

Figure 3 presents the Factual Recall and Hallucination Recall of various LLMs on the PerHalluEval Benchmark for QA. Key findings are explained below.

Hallucination Recall remains challenging. Most models struggle to identify hallucinated answers accurately. For example, Gemini 2.0 Flash achieves a Hallucination Recall of only 43.4% (95% CI: 41.8–45), while Llama 3.3 70B reaches 21.7% (95% CI: 20.4–23). This underscores how difficult a task it is—even for state-of-the-art models from major companies—to filter out hallucinations reliably.

Factual Recall and Hallucination Recall are often decoupled. High Factual Recall does not guarantee strong Hallucination Recall. Claude Sonnet 4 attains moderate Factual Recall (47.1%) but very high Hallucination Recall (81.3%), while Dorna2, despite its very high Factual Recall (96.7%), shows poor Hallucination Recall (4.4%).

Large variability among specialized Persian-tuned models. Persian-specialized models reach high Factual Recall (above 90%), but Hallucination Recall varies. PersianMind (Factual: 96.2% (95% CI: 95.6–96.8)) has low Hallucination Recall (under 6%), whereas Dorna2 exhibits similar trends, reflecting variability even among strong fine-tuned models on Persian.

Underperformance in smaller models. Smaller multilingual models like Qwen-2.5-7B display low Factual Recall (around 40%) and similarly weak Hallucination Recall, revealing the limitations of compact model architectures in both Factual and Hallucination Recall.

4.2. Summarization

Figure 4 presents the Factual Recall and Hallucination Recall of various LLMs on the PerHalluEval Benchmark for summarization. Key findings are explained below.

Models exhibit skepticism in accepting summaries. Several models tend to confidently reject hallucinated summaries (high Hallucination Recall) while being more conservative when accepting correct summaries (lower Factual Recall). For instance, Gemini 2.0 Flash-Lite demonstrates high Hallucination Recall (99.1% (95% CI: 98.7–99.3)) but lower Factual Recall in summary acceptance (65.5% (95% CI: 64–67)).

High overall Hallucination Recall. Most models perform strongly on summary hallucination detection. Top performers Qwen 2.5 7B (99.9%), LLaMA

3.1 8B (99.4%), Gemini 2.0 Flash-Lite (99.1%), and Claude Sonnet 4 (97.9%) demonstrate robust discrimination between faithful and hallucinated summaries.

Large variability among specialized Persian-tuned models. Persian-specific models demonstrate divergent results. Dorna achieves both high Factual Recall in summarization (95.8% (95% CI: 95.1–96.4)) and high Hallucination Recall (95.3% (95% CI: 94.6–95.9)). In contrast, PersianMind, despite robust Factual Recall (81.2% (95% CI: 79.9–82.4)), reaches only moderate Hallucination Recall (37.5% (95% CI: 36–39)), highlighting continued variability among targeted Persian-tuned models.

5. Discussion and Further Analysis

5.1. Persian vs. English QA: Minimal Performance Gap

To evaluate the internal knowledge of LLMs regarding Persian-specific concepts without any external context, this analysis is based on the QA task. To ensure the quality of the PerHalluEval question answering benchmark, a manual annotation process is conducted. Three annotators reviewed all samples with the guidelines explained in Appendix C.2, deciding whether the given context and questions were in a Persian context or not. The majority vote (at least two agreeing) decided the final label. Inter-annotator agreement, measured using Fleiss' kappa, yielded a high score of $\kappa = 0.87$, indicating strong reliability. Finally, 38% of the PerHalluEval QA benchmark is classified as Persian-specific.

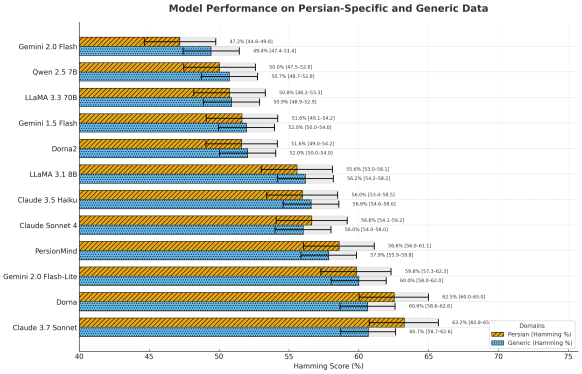


Figure 5: Comparison of LLM performance on Persian-specific vs. non-Persian content in the QA task (Hamming Score).

Model performance on Persian-specific and non-Persian content. The evaluation results are shown in Figure 5. Scores cluster in a relatively tight band (roughly mid-40s to low-60s), with modest differences between the two subsets. On the Persian-specific split, the strongest performance is delivered by Claude 3.7 Sonnet, followed closely by Dorna and Gemini 2.0 Flash-Lite. On the generic split, Dorna leads, with Claude 3.7 Sonnet and Gemini 2.0 Flash-Lite close behind. Lower-end performance is observed for Gemini 2.0 Flash and smaller/older open-source models such as Qwen-2.5-7B and LLaMA-3.3-70B.

Table 2: Comparison of evaluated LLMs on QA and text summarization (TS) using Hamming Score with 95% confidence intervals.

LLM Family	Model Name	Task	Hamming Score	95% CI
Anthropic	Claude Sonnet 4	QA	0.64	0.62–0.64
		TS	0.67	0.66–0.68
	Claude Sonnet 3.7	QA	0.63	0.61–0.63
		TS	0.81	0.80–0.82
	Claude Haiku 3.5	QA	0.57	0.55–0.57
		TS	0.82	0.81–0.83
Gemini	Gemini 1.5 Flash	QA	0.59	0.58–0.60
		TS	0.90	0.89–0.90
	Gemini 2.0 Flash-Lite-preview	QA	0.57	0.56–0.58
		TS	0.82	0.81–0.83
	Gemini 2.0 Flash	QA	0.62	0.60–0.62
		TS	0.91	0.90–0.91
Open-source	Qwen 2.5 7B	QA	0.47	0.45–0.48
		TS	0.53	0.52–0.54
	Llama 3.1 8B	QA	0.52	0.51–0.53
		TS	0.79	0.78–0.80
	Llama 3.3 70B	QA	0.57	0.55–0.58
		TS	0.83	0.82–0.83
	Dorna	QA	0.52	0.50–0.52
		TS	0.96	0.95–0.96
	Dorna2	QA	0.50	0.49–0.51
		TS	0.51	0.50–0.52
	PersianMind	QA	0.51	0.49–0.52
		TS	0.59	0.58–0.60

Performance gap between Persian-specific and non-Persian content. Directionally, the gap between Persian-specific and generic QA is small and sometimes mixed: a few models are slightly better on the Persian-specific items, while others favor the generic subset. The overlap of the 95% confidence intervals indicates that performance differences between Persian and English are not statistically significant and are likely not practically meaningful.

5.2. The Hamming Olympics: Where Summarization Takes the Gold!

The comparison between LLMs' performance on QA and summarization tasks is illustrated in Table 2. One of the most remarkable cross-task observations is that Hamming Scores on summarization are significantly higher than QA scores across almost all assessed models. This performance disparity can be interpreted in terms of intrinsic and extrinsic hallucinations. In question answering, models depend largely on their internal knowledge, and due to Persian being a low-resource language, the Hamming Score decreases when operating in the Persian context. Conversely, when evaluating LLMs in the summarization task, the original article is provided to the models. This external knowledge guides models to better decide whether the sum-

mary contains hallucinated content or not. Two open-source models, Qwen 2.5 7B and Dorna2, are exceptions as they perform well on QA as well as summarization. Models such as Claude Sonnet 3.7, Gemini 1.5 Flash, Gemini 2 Flash, and Llama 3.3 70B consistently perform at the top across both QA and summarization tasks.

5.3. The 'Yes-Man' Problem: Persian-Tuned LLMs Can't Say No

In the QA task, Persian-tuned models (Dorna, Dorna2, and PersianMind) obtained the three lowest performance scores. This observation can be explained by examining the results presented in Figure 3, in which these models demonstrate the highest Factual Recall. Specifically, the models consistently exhibit a strong inclination toward affirming that the presented answer or content is not hallucinated, regardless of its factual accuracy. Consequently, when assessed using Hallucination Recall, the Persian-tuned models achieve notably lower scores. This disparity indicates a bias within these models toward incorrectly classifying content as accurate and non-hallucinated, thereby resulting in diminished performance on tasks explicitly measuring hallucination sensitivity.

5.4. A Case Study: Summarization Performance Drop after Persian Fine-Tuning

Task-dependent behavior is demonstrated by a comparison between Dorna2 and its base model, Llama 3.1 8B. The models' Hamming Scores in QA are nearly the same, indicating that Persian fine-tuning had a small effect on this task for this particular model pair (see Table 2). In comparison, there is a noticeable difference in the summarization task. Figure 4 shows that Dorna2's Hallucinated Recall is about 60 percentage points lower than Llama 3.1 8B's. Additionally, the summarization's Hamming Score steadily declines from 0.79 (Llama 3.1 8B) to 0.51 (Dorna2). These findings suggest that in this particular case, summarization seems to be more sensitive to fine-tuning effects.

6. Conclusion

In this paper, we introduce PerHalluEval, the first benchmark developed specifically for evaluating hallucination in the Persian language. We propose a three-step automated pipeline to generate appropriate hallucinated samples systematically. In the first stage, suitable samples are extracted from well-known Persian datasets of QA and summarization tasks to analyse extrinsic and intrinsic hallucinations. Subsequently, in the generation stage, three distinct hallucinated candidate outputs are generated for each selected sample, harnessing three powerful LLMs and a strong instruction. The most believable hallucinated candidate is chosen as the final sample after we use a state-of-the-art language model as a judge and log probabilities to evaluate plausibility. Additionally, we invite annotators to specify whether each data point in the QA dataset is specifically related to Persian culture or not for additional investigations.

Our detailed experimental evaluations across 12 diverse LLMs, including commercial, open-source, and specifically fine-tuned models on Persian, show notable limitations in detecting and mitigating hallucination in the Persian language. Our suggested metrics—Hallucination Recall, Factual Recall, and Hamming Score—demonstrate the nuanced challenges these models encounter, highlighting the substantial scope for future advancement in Persian LLMs.

With PerHalluEval, we address a critical gap by introducing a specialized Persian hallucination benchmark to support future NLP research. We believe our benchmark will facilitate further work towards developing trustworthy and culturally-aligned language models for Persian.

7. Ethical Considerations

This work creates a benchmark by selecting 4,000 items from pre-existing Persian corpora for QA (PQuAD) and summarization (PN-Summary). In order to maintain anonymity and confidentiality, three independent undergraduate native Persian annotators received instruction and calibration before labeling; participation was informed and voluntary, and annotator inputs were processed without personal identification. High agreement and majority vote validation of every item supported dependability while reducing personal burden.

We openly share our findings by providing model identifiers and deterministic assessment parameters to aid in critical review and replication.

8. Bibliographical References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei-Bidgoli. 2023. Persianllama: Towards building first persian large language model. *ArXiv*, abs/2312.15713.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat,

Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Made laine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, An drey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *arXiv preprint arXiv:2410.22257*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Zouying Cao, Yifei Yang, and Hai Zhao. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. arXiv preprint arXiv:2407.10793.

Andrew S Cassidy, Guillaume Garreau, Jay Sivagnaname, Mike Grassi, Bernard Brezzo, John V Arthur, and Dharmendra S Modha. 2025. Mitigating hallucinations and omissions in llms

for invertible problems: An application to hardware logic design automation. *arXiv preprint arXiv:2512.03053*.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023. Factchd: Benchmarking fact-conflicting hallucination detection. In *International Joint Conference on Artificial Intelligence*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *ArXiv*, abs/2310.03368.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. *ArXiv*, abs/2010.02443.

Esin Durmus, He He, and Mona T. Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *ArXiv*, abs/2005.03754.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and D. Reitter. 2021. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE.

Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli SadraeiJavaheri, Shayan Bali, Amirhossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, et al. 2025a. Melac: Massive evaluation of large language models with alignment of culture in persian language. *arXiv preprint arXiv:2508.00673*.

Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri, and Saeedeh Momtazi. 2025b. Persian in a court: Benchmarking vlms in persian multi-modal tasks. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 52–56.

Farhan Farsi, Sadra Sabouri, Kian Kashfipour, Soroush Gooran, Hossein Sameti, and Ehsaneddin Asgari. 2024. Syntran-fa: Generating comprehensive answers for farsi qa pairs via syntactic transformation.

Arash Fayyazi, Mahdi Nazemi, Arya Fayyazi, and Massoud Pedram. 2024. Neuroblend: Towards low-power yet accurate neural network-based inference engine blending binary and fixed-point convolutions. In *Proceedings of the Great Lakes Symposium on VLSI 2024*, pages 730–735.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language?

Masood Ghayoomi, Saeedeh Momtazi, and Mahmood Bijankhan. 2010. A study of corpus development for persian. *Int. J. Asian Lang. Process.*, 20:17–34.

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, Navy-

ata 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.

Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello. 2024. Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks. *Scientific reports*, 14(1):16375.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43:1 – 55.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah: Analytical annotation of hallucinations in large language models. In *Proceedings of ACL 2024*.

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 computing surveys*, 55(12):1–38.

Safoura Aghadavoud Jolfaei and Azadeh Mohebi. 2025. A review on persian question answering systems: from traditional to modern approaches. *Artif. Intell. Rev.*, 58:127.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. *arXiv preprint arXiv:2403.14859*.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2020. Parsinlu: A suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *ArXiv*, abs/2406.07070.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.

Ali Mehrban and Pegah Ahadian. 2023. Evaluating bert and parsbert for analyzing persian advertisement data. *ArXiv*, abs/2305.02426.

Swaroop Mishra and Elnaz Nouri. 2023. Help me think: A simple prompting strategy for non-experts to create customized content with models.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Thomas Petersen, Pouya Golchin, Jinwoo Im, and Felipe PJ de Barros. 2025. Electrokinetic effects on flow and ion transport in charge-patterned corrugated nanochannels. *arXiv preprint arXiv:2510.22182*.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *ArXiv*, abs/2401.06466.

Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2022. naab: A ready-to-use plug-and-play corpus for farsi. *arXiv preprint arXiv:2208.13486*.

Seyed Mojtaba Sadjadi, Zeinab Rajabi, Leila Rabiei, and Mohammad-Shahram Moin. 2024. Farssibert: A novel transformer-based model for semantic similarity measurement of persian social networks informal texts. *ArXiv*, abs/2407.19173.

Mehrnoush Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *International conference language technologies for all (LT4All)*.

Faezeh Dehghan Tarzjani and Bhaskar Krishnamachari. 2025. Learning wireless interference patterns: Decoupled gnn for throughput prediction in heterogeneous multi-hop p-csma networks. *arXiv preprint arXiv:2510.14137*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *ArXiv*, abs/2004.04228.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024b. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Hanzhi Zhang, Sumera Anjum, Heng Fan, Weijian Zheng, Yan Huang, and Yunhe Feng. 2025. Polyfever: A multilingual fact verification benchmark for hallucination detection in large language models. *arXiv preprint arXiv:2503.16541*.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization.

A. Reproducibility and Implementation Details

A deterministic evaluation protocol was adopted. Unless precluded by provider defaults, all systems were evaluated with temperature $= 0.0$, top-$p = 1.0$, and $\texttt{max\_tokens} = 512$. Commercial models were accessed exclusively through their official APIs, and open-source models were executed locally on Google Colab under the indicated GPUs. All open-source models can be found in their respective Hugging Face repositories. Exact identifiers are enumerated in Table 3, which is intended to enable faithful replication of the reported results.

B. Hamming Score

In our model, each input pair generates two binary outputs—one "Factual" and one "Hallucinated"—encoded as a vector $y \in \{0,1\}^2$ with the Factual label $[1,0]$. Let $f : \mathcal{X} \to \mathbb{R}^2$ represent our scoring function and t be a thresholding operator (e.g., $t(f_j(x)) = 1$ if $f_j(x) \geq 0.5$ otherwise, 0). The classifier is defined as $H(x) = t(f(x))$.

We define the multi-label Hamming Loss and Hamming Score as:

$$L_H(H(x), y) = \frac{1}{2} \sum_{j=1}^{2} [\![t(f_j(x)) \neq y_j]\!],$$

$$HammingScore = 1 - L_H(H(x), y),$$

Model	Access	Identifier / Repo	Env./API	T	top-p	Max
GPT-4o	Closed (API)	`gpt-4.1-2025-04-14`	OpenAI API	0.0	1.0	512
Claude Sonnet 4	Closed (API)	`claude-sonnet-4-20250514`	Anthropic API	0.0	1.0	512
Claude Sonnet 3.7	Closed (API)	`claude-3-7-sonnet-20250219`	Anthropic API	0.0	1.0	512
Claude Haiku 3.5	Closed (API)	`claude-3-5-haiku-20241022`	Anthropic API	0.0	1.0	512
Gemini 1.5 Flash	Closed (API)	`gemini-1.5-flash`	Google AI Studio	0.0	1.0	512
Gemini 2.0 Flash	Closed (API)	`gemini-2.0-flash`	Google AI Studio	0.0	1.0	512
Gemini 2.0 Flash-lite	Closed (API)	`gemini-2.0-flash-lite`	Google AI Studio	0.0	1.0	512
LLaMA 3.1 8B Instruct	Open (Local)	`meta-llama/Llama-3.1-8B-Instruct`	Colab L4	0.0	1.0	512
LLaMA 3.3 70B Instruct	Open (Local)	`meta-llama/Llama-3.3-70B-Instruct`	Colab A100	0.0	1.0	512
Qwen 2.5 7B Instruct	Open (Local)	`Qwen/Qwen2.5-7B-Instruct`	Colab L4	0.0	1.0	512
Dorna	Open (Local)	`PartAI/Dorna-Llama3-8B-Instruct`	Colab L4	0.0	1.0	512
Dorna2	Open (Local)	`PartAI/Dorna2-Llama3.1-8B-Instruct`	Colab L4	0.0	1.0	512
PersianMind	Open (Local)	`universitytehran/PersianMind-v1.0`	Colab L4	0.0	1.0	512

Table 3: Evaluation configuration. Deterministic decoding was used unless a provider's default prevented overriding a parameter. "Max" denotes `max_tokens`.

C. Annotation Guidelines

C.1. General Guideline for Human Validation

Detailed Annotation Guideline

You will be shown pairs of text: a gold standard correct version and a potentially hallucinated version given by a model. Your task is to determine whether the model's version comprises hallucination(s) in terms of the following types:

- **Contextual misinterpretation**: Misunderstanding the context of the source text and coming up with irrelevant or distorted work.
- **Factual fabrication**: Inclusion of matter not in the source and not verifiable.
- **Specificity distortion**: Making the text significantly more general or excessively specific than the source.
- **Wrong assumption**: Drawing a conclusion not warranted by the statement given.
- **Unnecessary doubt**: Suffering from unnecessary doubt or uncertainty about the right facts.

Mark the sample as **"Hallucinated"** if any of the above types are represented; otherwise, mark it as **"Factual."** If you are unsure, choose the label you consider best and leave a brief comment explaining your choice.

Focus on the meaning rather than grammar, fluency, or style. Wording differences that do not affect meaning should not be classified as hallucinations. If in doubt, re-read the source text carefully and verify that all significant facts remain valid and reliable in the model's answer.

Annotators were undergraduate native Persian speakers with no overlap with the authorship team. They received training examples for each hallucination type and participated in a calibration round prior to full annotation.

C.2. Annotating Guideline for Persian-specific and Generic Content

Annotators also received explicit instructions to distinguish Persian-specific from generic content.

Detailed Guideline for Persian-specific vs Generic Content

You will be shown Persian-language content. Your task is to determine whether it specifically relates to Persian culture or history, or if it is generic content without unique Persian aspects.

- Look for clear references to landmarks, cities, personalities, historical eras, or cultural expressions associated with countries where Persian is an official language (Iran, Afghanistan, Tajikistan).
- Examples of Persian-specific content include mentions of Persian cities, important historical figures, dynasties, traditional celebrations, or cultural institutions.
- If the content does not emphasize uniquely Persian elements, it should be labeled as generic.

Focus on cultural or historical specificity rather than language alone. Content written in Persian but describing non-Persian topics should not be considered Persian-specific.

D. Error analysis cases

This section presents four representative examples of hallucination errors identified in our evaluation dataset. Each entry includes the question (with English translation when originally in Persian), the hallucinated answer produced by models, the gold (ground-truth) answer, and an analysis of model behavior. These cases highlight different error types that remain challenging for current LLMs.

Factual Fabrication – Numeric Fact

Question:
In which year did **Andi Gutmans** and **Zeev Suraski** lay the foundations of **PHP 3**?

Hallucinated Answer:
1997 AD

Gold Answer:
1995 AD

Analysis:
Only 1 out of 12 evaluated models (Gemini 1.5 Flash) flagged this answer as incorrect. The other 11 accepted the fabricated date, revealing a shared weakness in recalling precise historical numbers.

Contextual Misinterpretation

Question:
What were the demonstrators in **Bukan** demanding in 1979?

Hallucinated Answer:
They were demanding the establishment of an Islamic Republic.

Gold Answer:
They were demanding the release of political prisoners from the Bukan region and its surroundings.

Analysis:
Only 3 models (assessed Gemini models) flagged this hallucinated answer as incorrect. The other 9—including all three Persian-tuned models—accepted it, indicating difficulty grounding answers in historical socio-political contexts.

Contextual Misinterpretation Driven by an Ezāfe Noun Chain

Question:
Who collaborated on the restoration of the film **The Splendor of Persepolis**?

Hallucinated Answer:
Japanese and Italian experts.

Gold Answer:
A group of Iranology scholars from leading universities worldwide.

Analysis:
Only 2 models (Gemini 1.5 Flash and Claude Sonnet 3.7) answered correctly. The question included a four-word Persian ezāfe noun chain encoding a single proper name ("The Splendor of Persepolis film"). Most models misparsed it, treating only "restoration" as the head and thus accepting a generic hallucinated answer.

Incorrect Inference

Question:
How did **Shah Abbas** drive the Portuguese out of **Bandar Abbas**?

Hallucinated Answer:
By launching a military assault and occupying the port.

Gold Answer:
By engineering commercial rivalry among European powers on the Persian Gulf coast.

Analysis:
Only Gemini 2.0 Flash detected the hallucinated answer. Although the supporting facts were present in the context, the hallucinated response leapt to an unsupported causal inference (direct warfare) instead of the documented indirect economic strategy described in the gold answer.

E. Filtering Prompts

Figures 6 and 7 are used in the filtering stage to select the best candidate.

F. Precision–Recall Analysis with Iso-F1 Curves

Insights into model behavior beyond the single scalar metric were sought by plotting the trade-off between precision and recall for all of the models across both tasks QA and summarization (Figure 8). Each point was represented as a model, with its position determined by its precision on the x-axis and its recall on the y-axis. Bubble sizes were scaled proportionally to the F1 scores of the respective models, and dotted iso-F1 curves were included as visual guides to aid in the interpretation of the precision/recall trade-off. Through these graphs (Figure 8a,b), models exhibiting high recall with minimal loss of precision, as well as those biased toward either precision or recall, were revealed.

G. Log Probabilities Regex

The regular expression for 'Y' is:

```
[ yYỳỲýÝÿŸȳȲẏẎŷŶ¥'.-]+
```

The regular expression for 'N' is:

```
[ nNñÑńŃňŇṇṆn̈N̈'.-]+
```

H. Summary of Best Results of Each Model Family

Table H is the table containing a summary of results for QA and summarization with three metrics in each model family.

I. Hallucinated content generation prompts

The full prompt for generating hallucinated answers for task QA can be found in Figure 9. The full prompt to generate hallucinated summaries can be found in Figure 10.

Summarization Task

Prompt:
You are a text summary evaluation system. Your task is to accurately compare the provided summary with the original text. When evaluating, focus only on these criteria:

1. Has all the information in the summary been accurately extracted from the original text?
2. Is there any incorrect, distorted, or additional information in the summary?

Input:
Article:
Clarence Seedorf, the Dutch football legend, recently visited Iran for a leisure trip. During his stay, he met with several prominent figures in Iranian football, including Ali Daei, Karim Bagheri, Vahid Hashemian, and the popular sports journalist Adel Ferdosipour. According to the latest reports from IMNA, Ferdosipour conducted an exclusive interview with Seedorf. This interview is set to be broadcast on the "Football 120" program, which airs Thursday night at 11 PM on Iran's Varzesh (Sports) TV channel.

Summary:
"The producer of Football 120 has held talks with the officials of the Varzesh (Sports) channel."

Output:
"N"

Log Probabilities:

```
{ "N":  -0.2357, "Ń":  -4.6052,
"Y":  -1.9661, "Ň":  -4.6052,
"ń":  -2.9957 }
```


Question Answering Task

Prompt:
You are a response validity checking system. Your task is to verify whether the provided answer to a specific question is correct or not. You are only allowed to provide one of two letters: 'Y' (for correct answer) or 'N' (for incorrect answer) as the final output. Do not provide any explanations, extra text, or conclusions. Only give one letter as output: Y or N.

Input:
Question:
What place was called Persepolis or Parseh?

Answer:
Persepolis or Parseh is the name of one of the ancient cities of Greece that was known as a cultural and artistic center during the Achaemenid Empire.

Output:
"N"

Log Probabilities:

```
{ "N":  -0.3711, "Ñ":  -4.6052,
"n":  -1.8971, "Y":  -2.9957,
"ŋ":  -2.3026 }
```


Figure 6: Instructions of hallucination filtering for question answering and summarization tasks. The Figure shows the English translation of the original Persian prompts.

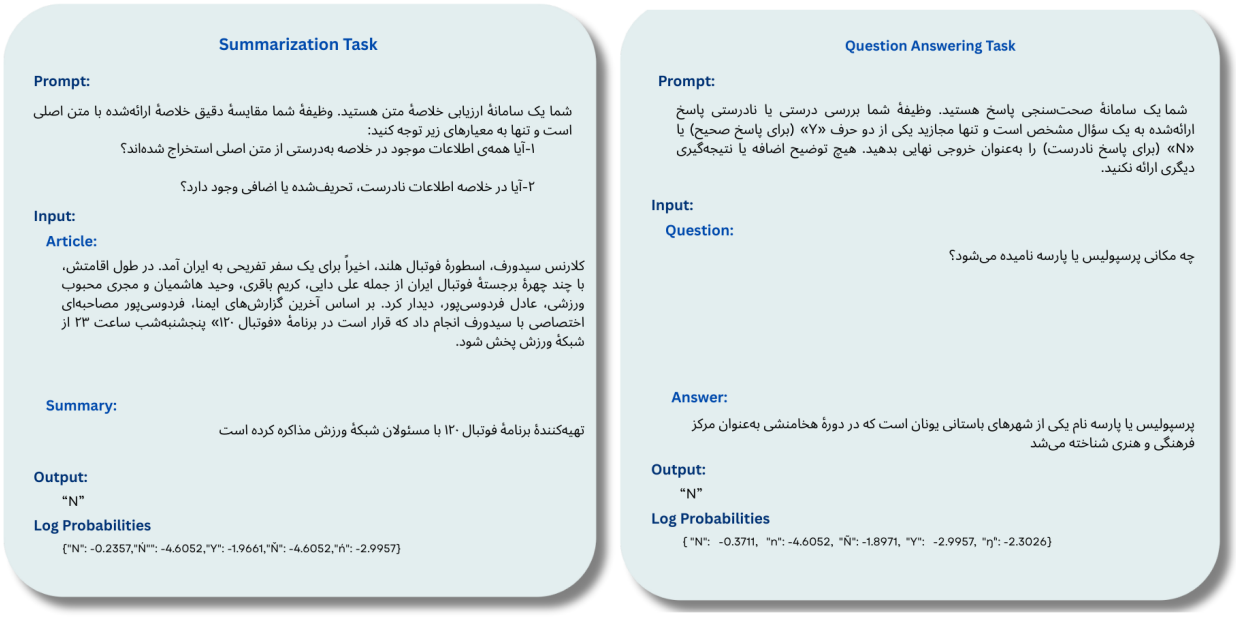
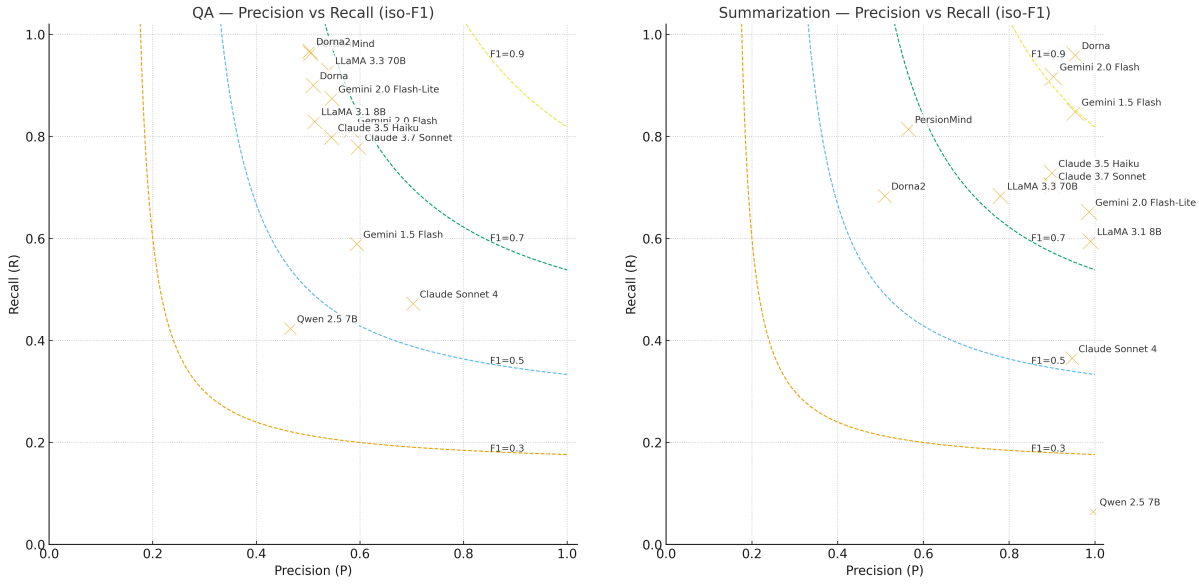


Figure 7: Hallucination filtering guidelines for question-answering and summarization tasks. The image displays the original prompts in Persian.



(a) QA Task (b) Summarization Task

Figure 8: Precision–Recall maps with iso-F1 curves. Each point is a model, bubble size reflects its F1 score, and iso-F1 curves (gray) provide guidance on the precision–recall trade-off.

Family	Task	Metric	Best variant	Result
Anthropic	QA	Hamming Score	Claude Sonnet 4	0.64
		Factual Recall	Claude Haiku 3.5	80.0%
		Hallucinated Recall	Claude Sonnet 4	81.3%
	TS	Hamming Score	Claude Haiku 3.5	0.82
		Factual Recall	Claude Haiku 3.5	72.9%
		Hallucinated Recall	Claude Sonnet 4	97.9%
Gemini	QA	Hamming Score	Gemini 2.0 Flash	0.61
		Factual Recall	Gemini 2.0 Flash-lite	87.5%
		Hallucinated Recall	Gemini 1.5 Flash	61.0%
	TS	Hamming Score	Gemini 2.0 Flash / Gemini 1.5 Flash	0.90
		Factual Recall	Gemini 2.0 Flash	91.8%
		Hallucinated Recall	Gemini 2.0 Flash-lite	99.1%
Fine-tuned on Persian	QA	Hamming Score	Dorna	0.51
		Factual Recall	Dorna2	96.8%
		Hallucinated Recall	Dorna	13.7%
	TS	Hamming Score	Dorna	0.95
		Factual Recall	Dorna	95.8%
		Hallucinated Recall	Dorna	95.3%
Open-Source Multilingual models	QA	Hamming Score	Llama 3.3 70B	0.56
		Factual Recall	Llama 3.3 70B	93.0%
		Hallucinated Recall	Qwen 2.5 7B	52.3%
	TS	Hamming Score	Llama 3.1 8B / Llama 3.3 70B	0.79
		Factual Recall	Llama 3.3 70B	68.3%
		Hallucinated Recall	Qwen 2.5 7B	99.9%

Table 4: Comparison of model families across QA and summarization tasks. This table shows the best-performing models within each family based on different evaluation metrics (Hamming, Factual Recall, and Hallucinated Recall). Model families include Anthropic, OpenAI, Gemini, fine-tuned Persian models, and open-source multilingual models.

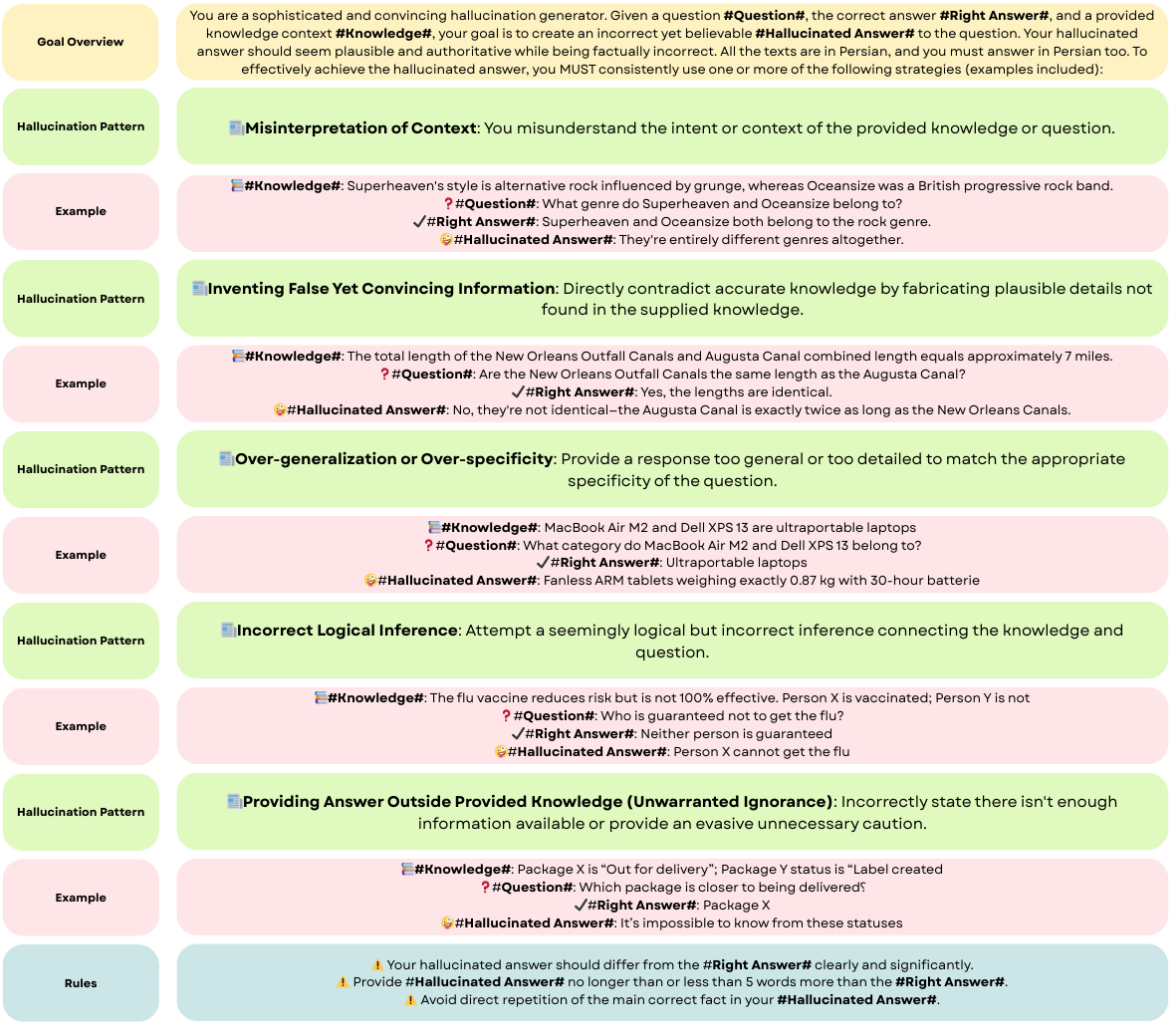


Figure 9: The prompt used for QA hallucinated content

Figure 10: The prompt used for Summarization hallucinated content