TruthTrap: A Bilingual Benchmark for Evaluating Factually Correct Yet Misleading Information in Question Answering

Mohammadamin Shafiei[*1,4], **Hamidreza Saffari**[*2],
Mohammad Taher Pilehvar[3], **Alessandro Raganato**[4]
[1]University of Milan, [2]Politecnico di Milano,
[3]Cardiff University, [4]University of Milano-Bicocca
m.shafieiapoorvari@studenti.unimi.it hamidreza.saffari@mail.polimi.it
pilehvarmt@cardiff.ac.uk alessandro.raganato@unimib.it

Abstract

Large Language Models (LLMs) are increasingly used to answer factual, information-seeking questions (ISQs). While prior work often focuses on false misleading information, little attention has been paid to true but strategically persuasive content that can derail a model's reasoning. To address this gap, we introduce a new evaluation dataset, TRUTHTRAP, in two languages, i.e., English and Farsi, on Iran-related ISQs, each paired with a correct explanation and a true-yet-misleading hint. We then evaluate nine diverse LLMs (spanning proprietary and open-source systems) via factuality classification and multiple-choice QA tasks, finding that accuracy drops by 25%, on average, when models encounter these misleading yet factual hints. Also, the models' predictions match the hint-aligned options up to 77 percent of the time. Notably, models often misjudge such hints in isolation yet still integrate them into final answers. Our results highlight a significant limitation in LLM outputs, underscoring the importance of robust fact-verification and emphasizing real-world risks posed by partial truths in domains like social media, education, and policy-making. Our dataset is openly available at https://github.com/Mamin78/truthtrap_with_code.

1 Introduction

The adoption of Large Language Models (LLMs) in NLP has brought considerable improvements in several tasks, spanning generation and classification (Zhao et al., 2023; Hurst et al., 2024). These models excel at generating human-like text (Liu et al., 2023; de Souza et al., 2025; Perchik, 2023; Wang et al., 2023) and have been harnessed for a range of applications, with question answering (QA) among others (Zhuang et al., 2023; Monteiro et al., 2024; Lucas et al., 2024; Arefeen et al., 2024). A central challenge within QA involves information-seeking questions (ISQs) (Saracevic et al., 1988), which require precise retrieval or inference from textual sources. These ISQs appear in critical contexts such as healthcare, education, and policy-making (Eskola, 1998; Limberg and Sundin, 2006; van Lieshout et al., 2020; Scacco and Muddiman, 2020; Mishra et al., 2015).

While common QA benchmarks such as SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), TyDi QA (Clark et al., 2020), and MLQA (Lewis et al., 2020) primarily assess answer accuracy given explicit textual evidence, they typically do not address the phenomenon of true but non-answer content, statements that appear pertinent yet fail to resolve the actual query. Figure 1 shows an example of how a distractor can overshadow the correct answer.

Such "true, persuasive" snippets can be especially misleading (Jin et al., 2024; Li et al., 2024; Saenger et al., 2024), particularly for instruction-tuned models that typically regard accurate text as reliable, even if it distracts from the correct answer. Previous research in adversarial or misleading QA has often examined false hints or conflicting contexts (Jia and Liang, 2017; Liu et al., 2025), while less attention has been paid to true-but-persuasive content. Related efforts in misinformation detection, such as FEVER (Thorne et al., 2018) and HoVer (Jiang et al., 2020), investigate claim verification against relevant documents but do not examine how factually correct yet irrelevant snippets can mislead the QA process.

To investigate how factually correct but distracting details can mislead LLMs, we introduce a new bilingual dataset in English and Farsi, where each information-seeking question contains a correct explanation and a persuasive-but-off-target hint. We curated and manually verified 1,000 multiple-choice QA items across ten categories; each item provides four answer choices, a truthful explanation aligned with the correct answer, and a mislead-

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EACL 2026, pages 2966–2987
March 24-29, 2026

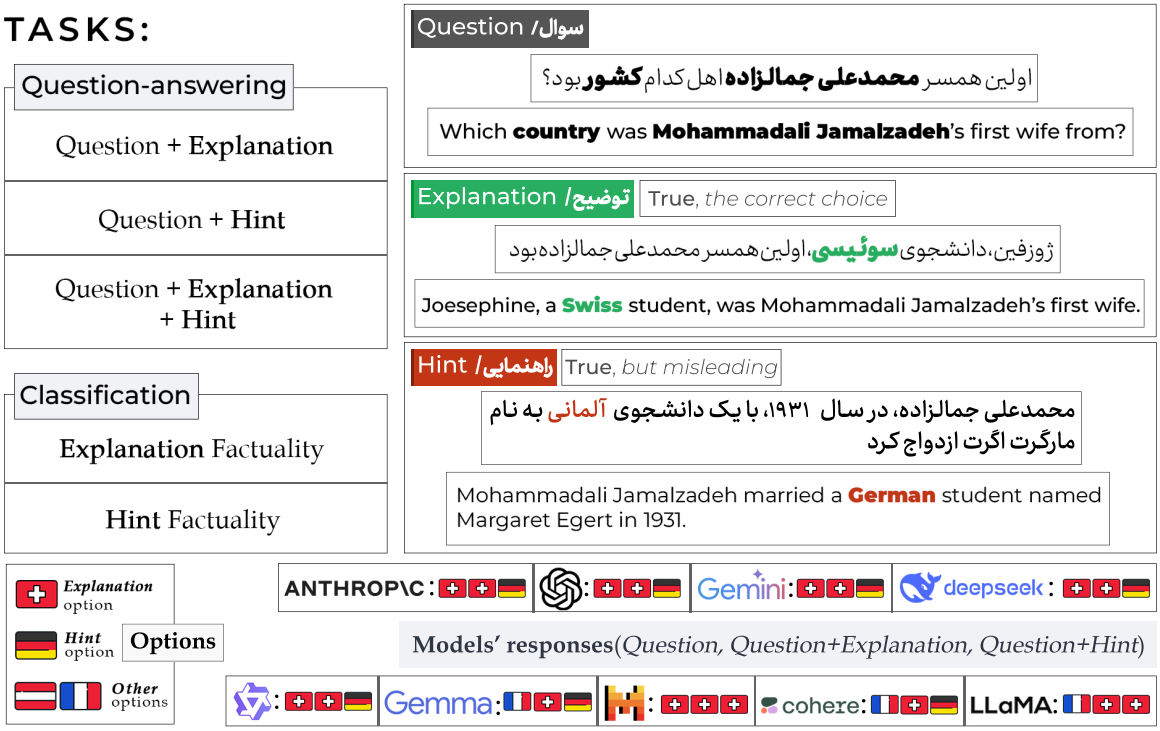

Figure 1: A bilingual QA example from the "Art and Literature" category, showing both English and Farsi question texts ("Which country was Mohammad Ali Jamalzadeh's first wife from?") and four candidate answers (Switzerland, Germany, Austria, France). The correct answer, Switzerland, is supplied by the Explanation, which identifies Josephine, a Swiss student whom Jamalzadeh married in 1914. However, the Persuasive Hint instead presents a factually accurate account of his second marriage in 1931 to Margaret Egert, a German student in Geneva. Although true, this detail is irrelevant to the first-wife question and can induce the model to choose Germany, showing how true but distracting information may override the intended answer. We test simple QA, QA with explanation, as well as QA with hints. The figure also shows the models' responses to this sample under the three settings. As shown, providing an explanation consistently leads models to answer Switzerland. In the hint-only setting, most models rely on the hint, except for Mistral and LLaMA.

ing hint that, although factually accurate, fails to address the actual query. Our experiments focus on five scenarios: QA without any added context, QA with the correct explanation, QA with the misleading hint, factuality classification of explanations, and factuality classification of hints. We tested recent LLMs such as GPT-4o, Claude, LLaMA, DeepSeek, and Qwen, finding that accuracy drops by 25%, on average, in the presence of true-but-distracting hints, with cross-lingual performance differences sometimes around 10%. Our error analysis shows that the most challenging cases arise when a hint partially overlaps with the correct explanation. In such situations, models often treat both as simultaneously valid and are consequently misled into selecting an off-target detail. By "partial overlap," we refer to cases in which the hint and the explanation share entities or partial facts (such as the same person, institution, or year) but differ in the critical information required to resolve the question correctly.

This paper provides the following contributions: (1) A novel bilingual dataset specifically designed to study the impact of true-but-misleading on information-seeking QA tasks, addressing a notable gap in adversarial QA literature, which typically focuses only on false or fabricated statements. (2) Multilingual comparative analysis involving English and Farsi, languages with contrasting resource availability, to explore how linguistic differences influence the susceptibility of models to factually correct but irrelevant information. (3) Systematic evaluation of several state-of-the-art LLMs, investigating their behavior when encountering factual yet misleading snippets, including error analysis. (4) One of the first large-scale Farsi resources aimed at evaluating recent LLMs, providing a challenging benchmark for model performance assessment in a lower-resource language setting, especially on ISQs.

2 Related Work

2.1 Information-seeking Questions

ISQs drive a large part of QA research. The paradigm started by focusing on building datasets that reflect the complexity of real-world queries (Narayanan et al., 1999; Ofran et al., 2012; Park et al., 2021; Shafiei et al., 2025). Dasigi et al. (2021) introduced questions based on full-text papers, requiring models to pull information from multiple sections. Similarly, Asai and Choi (2021) highlighted multilingual QA challenges, such as answerability and paragraph retrieval across languages.

With the rise of LLMs, research has shifted toward how well LLMs handle complex queries. For instance, Pang et al. (2024) examined LLMs' ability to interpret tables while Kamalloo et al. (2023) developed a dataset to support generative models that explain their answers with both human and machine-generated input. Moreover, ISQs explore various domains and categories, such as medicine, history, education (Golany et al., 2024; Chowdhury and Chowdhury, 2024; Fernández-Pichel et al., 2024; Yun and Bickmore; Mannuru et al., 2024).

In Farsi, however, existing ISQ resources are relatively few and often straightforward, allowing LLMs to perform at high accuracy. Khashabi et al. (2021) created a widely used Farsi QA benchmark that includes ISQs, but many items are quite basic (e.g., "What is the largest continent?"). Other Farsi QA datasets (Ghahroodi et al., 2024; Emami and Mosharraf, 2023) do not necessarily focus on ISQs or may be limited to school-level queries, thus not mirroring real-world difficulties. Still other works investigate the intersection of QA with social norms or cultural values in Farsi (Moosavi Monazzah et al., 2025; Saffari et al., 2025b,a; Sadr et al., 2025; Bokaei et al., 2025; Chang et al., 2025), but they differ in scope from the questions we pose about misleading yet factually correct content.

2.2 LLM Persuasion and Effects of Extra Context

LLMs are notably susceptible to various forms of persuasion, a topic that has been widely studied across different domains (Zeng et al., 2024; Rogiers et al., 2024). Some works have examined how personas affect model persuasion (Salewski et al., 2023; de Araujo and Roth, 2024), while others have focused on the effects of framing—especially emotional framing—in contexts like political discourse and health messaging (Jeng et al., 2024). Still others have compared persuasive responses between humans and LLMs (Carrasco-Farre, 2024).

In addition, previous research has shown that LLM outputs may shift upon receiving added context, even when such context is irrelevant (Anagnostidis and Bulian, 2024; Shi et al., 2023; Vishwanath et al., 2025). Nonetheless, little attention has been given to situations where factually correct yet off-target information disrupts LLM output, particularly for ISQs. Our new resource, TRUTHTRAP, is designed to fill this gap by assessing how true-but-misleading content impacts QA performance, offering insights into the broader phenomenon of LLM susceptibility to distractors that, while accurate, fail to resolve the actual query.

3 Dataset

In this section, we present our bilingual (Farsi-English) dataset, which targets true-but-misleading information in information-seeking questions (ISQs). Section 3.1 describes each QA item's composition, including the question, correct explanation, persuasive hint, and multiple-choice answers, all grounded in Wikipedia. Section 3.2 explains how we initially generated a pool of Iran-related ISQs using automatic methods and well-defined selection criteria. Finally, Section 3.3 details our human-driven curation and annotation pipeline, which refined the initial samples into 1,000 high-quality entries.

3.1 Dataset Framework

Each instance in our dataset comprises several elements that collectively shape our information-seeking QA task. All items are provided in both Farsi and English. Below is an overview of each component:

Question. An information-seeking prompt that queries specific factual details. For example, "When was the Institute of Journalism at the University of Tehran established, and with the help of which university?" The same question is provided in both Farsi and English.

Explanation. A correct statement, again in both languages, that supports the correct answer. Continuing the example above: "The Institute of Journalism at the University of Tehran was established in 1337 Solar Hijri (Iranian Calendar) with help from the University of Virginia, United States."

This explanation offers direct evidence for the right choice, ideally guiding the model to the accurate response.

Persuasive Hint. A truthful but irrelevant snippet meant to distract the model, typically on the same topic but not answering the actual question. In our example, the hint might read: "The University of Tehran, with help from Johns Hopkins University in 1343 Solar Hijri, established the doctoral program in cytopathology." While factually correct, this detail can misdirect the model into selecting the wrong answer.

Answer Options. Four potential answers, each plausible yet only one is correct. For the above question, these might be: (i) 1337 – Virginia University (correct), (ii) 1343 – Johns Hopkins University (related to the hint), (iii) 1333 – California University, (iv) 1320 – Utah University.

While only one persuasive hint is explicitly provided (in this case targeting the Johns Hopkins option), all distractors draw on true details relevant to the broader context. For instance, the California option reflects how the Institute of Administrative Sciences at the University of Tehran was developed in 1333 with assistance from the University of Southern California, and the Utah option references an earlier collaboration with Utah State University to expand agricultural programs. Even though these facts do not appear as separate persuasive hints, they maintain realism by offering multiple verifiable but off-target possibilities.

Categories and Subcategories. The dataset covers ten diverse categories: Arts & Literature, Education, Entertainment, Food, Geography, History, Holidays & Leisure, Religion, Science & Technology, and Sports. Each major category is further divided into three subcategories, yielding 100 questions per category (1,000 total). Subcategory definitions and their statistics are detailed in Appendix A.1. This breadth ensures a robust assessment of model behavior across various topics.

Question Type. Following Yang et al. (2018), each entry is labeled by the nature of the sought information, such as person, place, time, event, or artwork. Appendix A.1 details the distribution of these classes, highlighting the variety of ISQs in our dataset.

Target Type. Hints may undermine the correct answer generally (often in negatively phrased prompts) or reinforce a specific incorrect choice. For instance, in the question "Which dynasty did not choose Shiraz as the capital of Iran?", the hint might emphasize the Fars region (that includes Shiraz) without clarifying that the Sassanids used Estakhr (another city in Fars), not Shiraz, thus subtly misleading the model. Analyzing how models handle such general vs. targeted hints, like Figure 1, offers insight into potential adversarial vulnerabilities. Detailed statistics are provided in Appendix A.1.

Wikipedia Grounding. Each question links to at least one relevant Wikipedia page, ensuring all content, including explanations, hints, and distractors, is rooted in verifiable facts rather than fabricated statements. We deliberately selected Farsi Wikipedia pages during dataset construction: Farsi entries are typically richer and more detailed for Iran-related topics. A page is considered suitable if it (i) exists in Farsi, (ii) matches the intended subcategory, and (iii) contains substantive content rather than only metadata or stubs.

3.2 Initial Dataset Generation

To construct our initial, automatically generated sample of questions, we followed a two-stage process. First, we searched for suitable Wikipedia pages within each subcategory; second, we automatically created information-seeking questions (ISQs) using a few-shot prompt. We provide further details in Appendix A.2.

We began by using ChatGPT-4o's search capabilities to retrieve five relevant Wikipedia pages per subcategory, ensuring that each link led to a valid article. If a suitable page was missing, we iteratively requested additional candidates until we reached 15 pages per major category, yielding a total of 150 pages across the ten categories.

After compiling these pages, we automatically generated ISQs, along with explanations, hints, and multiple-choice options, via a few-shot prompt in Farsi using Claude Sonnet 3.7. For each Wikipedia page, this method produced 15 structured QA samples, resulting in 2,250 initial questions (15×150). These items served as a preliminary dataset, later refined through careful curation and annotation (Section 3.3).

3.3 Sample Selection and Annotation

We followed a multi-step annotation protocol to refine our initial large pool of automatically gener-

ated questions into a final set of 1,000 items (100 per category). Two annotators, both native Farsi speakers with backgrounds in Iranian studies and fluent in English, oversaw every stage of this process.

First, each question was evaluated to confirm whether it posed a valid information-seeking question (ISQ). For instance, "What is the average age of people who smoke in Iran?" requests factual data and is advanced to further review, whereas "Is smoking acceptable?" elicits an opinion and was removed. Only items passing this ISQ check proceeded to the next step.

Next, each retained question was examined for the accuracy of its explanation and proposed correct answer. The annotator carefully compared these elements with the relevant Wikipedia source. Any discrepancies led to revisions based on Wikipedia, or, if insufficient information existed, to discarding the question entirely. This procedure ensured that each explanation remained grounded in verified facts.

The third step focused on persuasive-but-truthful hints. When an automatically generated hint proved too vague or insufficiently misleading, such as "South Khorasan province is a large province in Iran and has a huge population, so it must have a large saffron cultivation area." for the query "Which province of Iran has a larger share in terms of saffron cultivation area?", the annotator refined or replaced it with more precise statements drawn from the same Wikipedia entry. For this specific example, the new hint is "The birthplace of saffron is Qaen county in South Khorasan province in Iran.". This new hint, compared to the automatically generated hint, provides an objective and sufficiently misleading fact. Items that could not accommodate a suitable hint were removed.

We then validated the other answer choices. Beyond the correct answer and the hint-based distractor, two additional options were chosen or revised to ensure they were accurate but incorrect for the actual query. If no suitable related facts were found in Wikipedia, the automatically generated options were retained if they were incorrect yet tangential to the question. Each item received a label for its question type (e.g., person, time, location) and hint target type (general misdirection vs. reinforcing a specific incorrect option). Once the first annotator completed this multi-step protocol and finalized the 1,000 items, the second annotator performed an independent review, confirming the correctness of explanations, hints, and distractors. Across all items, only 23 disagreements arose concerning question-type labels, often when one annotator used a more specific category (e.g., date/time) while the other used a broader one (e.g., proper/common noun). These discrepancies were resolved via brief discussion, favoring the finer-grained classification. No other issues emerged. Finally, we translated the curated Farsi dataset into English using Claude Sonnet 3.7. The same two annotators then applied a similar protocol to validate translations: the first ensured each English version was fluent, accurate, and faithful to the source, editing the translation when needed, while the second confirmed that no ambiguities remained.

4 Experimental Setup

This section describes how we use our bilingual (Farsi–English) dataset to investigate how additional context, whether an explanation or a true-but-misleading hint, affects LLMs in information-seeking questions. We conduct five experiments in both Farsi and English, comparing model outputs across different conditions. Table 1 shows the English prompt templates used in these experiments.

Task Design. Our experiments revolve around two main tasks: multiple-choice QA and factuality classification. In the multiple-choice QA task, models receive a question (Q) plus four answer options (A1, A2, A3, A4). We vary the presence of additional information to measure how explanations or hints guide or mislead the model:

- **Baseline QA (No Extra Context).** The model sees only Q and A1–A4, establishing a reference for accuracy without further details.
- **QA + Explanation.** The model is presented with Q, A1–A4, and a correct explanation that aligns with the right answer. This setup tests whether providing factual support boosts performance.
- **QA + Hint.** The model again sees Q and A1–A4 but now includes a true yet misleading hint. This setup tests whether introducing extraneous but accurate information degrades performance relative to the baseline.[1]

Beyond multiple-choice QA, we also run two

[1]We also tested a combined setup (**QA + Explanation + Hint**) and present those results in Appendix B. To isolate the impact of each information source, we focus in the main text on **QA + Explanation** or **QA + Hint**, highlighting how each independently influences model performance.

Mode	Prompt
Baseline	[Question] | 1: [First option] 2: [Second option] 3: [Third option] 4: [Fourth option] ONLY RETURN THE ANSWER OPTION'S NUMBER.
QA with additional information	[Question] | 1: [First option] 2: [Second option] 3: [Third option] 4: [Fourth option] Here is a piece of information: [Hint or Explanation] ONLY RETURN THE ANSWER OPTION'S NUMBER.
Factuality	Is this statement factually true, false, or are you uncertain and cannot determine for sure? "[Hint or Explanation]" ONLY RETURN ONE WORD FROM ['true', 'false', 'uncertain'] WITHOUT ANY KIND OF EXPLANATION.

Table 1: English prompt templates for the five experimental modes are shown below. The Farsi templates follow the same structure. The first row shows the baseline case, where questions are asked without any additional information. The second row shows prompts that include extra information alongside the question, either an explanation or a hint. The third row presents the classification prompt. In this factuality classification prompt, the additional information, either a hint or an explanation, is included within the prompt in the [Hint or Explanation] part. Together, these define five different test settings.

Model	Type	Farsi			English		
		T	F	U	T	F	U
Claude	exp	39.3	17.4	43.3	21.4	5.3	73.3
	hint	45.1	16.4	38.5	28.6	5.2	66.2
GPT	exp	57.2	15.0	27.8	49.4	25.3	24.8
	hint	59.0	17.1	23.9	53.1	23.9	21.9
Gemini	exp	37.6	45.0	17.4	34.6	35.4	30.0
	hint	42.4	41.2	16.4	41.4	31.6	27.0
DeepSeek	exp	57.1	8.2	34.6	24.8	5.5	69.7
	hint	66.3	5.6	28.1	35.5	4.7	59.8
Qwen	exp	22.0	53.8	23.6	20.9	22.6	56.2
	hint	30.9	45.3	23.4	28.4	18.8	52.1
Gemma	exp	59.5	15.2	25.3	57.7	7.6	34.7
	hint	65.2	12.7	22.1	66.5	6.4	26.7
Mistral	exp	4.9	0.0	95.1	6.3	3.1	90.6
	hint	7.2	0.0	92.8	11.4	3.4	85.2
Command	exp	80.4	10.0	9.6	14.3	1.5	84.2
	hint	79.3	5.7	15.0	24.1	0.9	75.0
LLaMA	exp	55.6	43.1	0.0	73.9	21.2	4.9
	hint	55.3	43.5	0.0	77.4	16.4	6.2
Average	exp	46.0	23.1	30.9	33.7	14.7	51.7
	hint	50.2	23.6	26.2	45.1	12.4	42.3

Table 2: Factuality classification accuracy results for explanations (exp) and hints (hint) in Farsi and English. Each model's outputs are split into proportions of True (T), False (F), and Uncertain (U) labels. The last row shows the averages across all models.

factuality classification experiments to assess whether the models can correctly label individual statements as True, False, or Uncertain:

- **Factuality of Explanations.** The model is given a factually correct explanation and asked to determine whether it is True, False, or Uncertain.
- **Factuality of Hints.** The model is provided with a factually correct hint and asked to classify it as True, False, or Uncertain.

These classification tasks clarify how well models identify truth in isolation, enabling us to distinguish between (a) failing to recognize a statement as factually true and (b) using that same statement incorrectly in QA. By construction, all explanations and hints in TRUTHTRAP are factually correct as they are extracted from and checked against Wikipedia, so the gold label is always True.

Models. We evaluated nine multilingual LLMs, covering both proprietary and open-source options: **Claude Sonnet 3.7**,[2] **GPT-4o** (Hurst et al., 2024), **Gemini-2.5-flash**, **Qwen3-next-80b-a3b-instruct** (Yang et al., 2024), **Llama-3.1-8b-instruct** (Meta et al., 2024)[3], **DeepSeek-V3** (Bi et al., 2024), **Mistral-small-3.1-24b-instruct**, **Command-r7b-12-2024**, and **Gemma-3-27b-it**. All experiments were conducted with a temperature setting of zero to ensure deterministic outputs. In the QA tasks, models were prompted to choose exactly one among A1–A4; in the factuality classification tasks, they were prompted to categorize a single statement as True, False, or Uncertain.

[2] https://www.anthropic.com/

[3] Even though LLaMa-3.1-8B do not officially support Farsi, recent works have shown its reliable performance on Farsi data, likely due to its extensive multilingual pretraining (Hosseinbeigi et al., 2025; Saffari et al., 2025b; Moosavi Monazzah et al., 2025; Zeinalipour et al., 2025).

Model	Language	Baseline (no context)		With Explanation		With Misleading Hints	
		ans_match	hint_match	ans_match	hint_match	ans_match	hint_match
Claude	Farsi	**62.16**	17.26	**99.10**	0.70	29.60	64.44
	English	**58.27**	19.80	**98.60**	1.30	24.54	67.65
GPT	Farsi	43.99	**41.58**	98.40	1.00	**40.82**	47.34
	English	57.40	17.95	98.20	1.30	**43.99**	41.58
Gemini	Farsi	60.76	16.22	98.80	1.00	34.44	55.49
	English	54.01	20.74	98.20	1.30	26.36	64.21
DeepSeek	Farsi	57.90	18.20	98.30	1.30	23.50	70.00
	English	53.00	21.10	98.50	1.30	17.03	**77.35**
Qwen3	Farsi	42.37	21.89	97.40	1.70	17.74	71.74
	English	46.73	22.36	97.80	1.50	21.93	66.10
Gemma	Farsi	49.85	21.51	97.60	1.70	22.12	65.66
	English	45.89	23.65	97.69	1.71	19.11	67.00
Mistral	Farsi	43.36	20.81	94.73	2.64	18.01	**72.40**
	English	44.58	23.30	97.49	1.71	18.83	68.62
LLaMA	Farsi	41.28	24.25	93.29	3.70	13.44	74.82
	English	36.03	25.90	92.86	**3.82**	15.53	70.34
Command	Farsi	35.78	25.23	86.65	**6.43**	16.63	68.44
	English	36.50	**29.50**	93.60	3.60	15.60	71.30
Average	Farsi	48.38	22.11	96.84	2.35	24.70	65.68
	English	48.27	22.70	96.77	1.95	22.77	67.13

Table 3: Comparison of model performance across three multiple-choice QA setups: **baseline with no context, with correct explanation provided**, and **with explicit misleading hints**. In each group, the column labeled "ans_match (accuracy)" indicates how frequently each system selects the correct answer, while "hint_match" shows how often it chooses the hint-based option.

5 Results and Discussion

We present our findings in two parts: factuality classification (Section 5.1), which evaluates how reliably models label individual factual explanations and hint statements as True, False, or Uncertain, and multiple-choice QA (Section 5.2), which explores how these statements, whether correct explanations or persuasive hints, affect QA accuracy. More results can be found in Appendix B.

5.1 Factuality Classification

Table 2 shows how each model labels explanations and hints (both factually correct) in Farsi and English; any choice other than True constitutes a classification error. There are different trends across the models. For instance, Claude, CommandR, and Deepseek perform much better in Farsi; LLaMA performs better in English; and some models show a relatively comparable performance across the languages, such as GPT, Gemini, Qwen, Gemma, and Mistral. Another observation is that models perceive False and Uncertain classes differently. For example, Claude, DeepSeek, Qwen, Gemma, and Mistral tend to answer with Uncertain a lot more than False. On the contrary, GPT, Gemini, and LLaMa answer with False and Uncertain labels similarly. CommandR shows a unique case. For Farsi, it shows a pattern like GPT, Gemini, and LLaMa, but for English, it is closer to what we see from the rest of the models. Among the models, LLaMa performs generally better in English compared to the rest of the models, reflecting its extensive post-training on factual data. In Farsi, CommandR is the one that performs better, which can be explained by their effort in handling multilingual data. Also, in general, models identify the explanations and hints more frequently as True in Farsi, even if the difference between Farsi and English rates is negligible. Overall, this discrepancy suggests that both domain specificity (in this case, Iran-related content) and multilingual alignment can shape how well a model identifies factual statements in different languages.

5.2 Multiple-choice QA

Baseline QA. In Table 3, the first super-column presents model performance on QA tasks with no additional context. Overall accuracy remains under 50%, averaging 48.27% in English and 48.38% in Farsi, underscoring the dataset's inherent difficulty. GPT, Gemini, DeepSeek, and Claude are the only models that achieve a score better than 50, at least for one language. Even Claude-3.7, a model

instrumental in constructing some of the dataset, only reaches about 60% accuracy. This suggests that despite leveraging Claude-3.7's capabilities in the dataset creation pipeline, our resource still poses a significant challenge to cutting-edge systems. Moreover, the open-source models, except for DeepSeek, do not even achieve an accuracy of 50%.

QA + Explanation. In Table 3, the second super-column shows model performance when given correct explanations aligned with the right answers. Accuracy often exceeds 90% in both Farsi and English, showing the strong impact of explicit factual support. Notably, models that rated explanations as Uncertain or False in isolation still use them effectively in the QA prompt. For instance, DeepSeek marks only 57.1% of Farsi explanations as True (and 34.6% as Uncertain), yet reaches over 90% QA accuracy with those same explanations. This effect is most striking for Mistral, which labels fewer than 10% of explanations as True but still benefits greatly in QA. Interestingly, while explanation judgments differ across languages for models like Claude, DeepSeek, CommandR, and LLaMa, these differences disappear in the QA setting. Overall, these findings confirm that explicitly correct context strongly boosts QA accuracy. However, some models appear more adept at integrating this context than others, illustrating that even if a model doubts an explanation in isolation, it often incorporates it as reliable once it is embedded in the prompt.

QA + Hint. The third super-column in Table 3 reports results when true-but-misleading hints are added to the QA setting. On average, hint-aligned answers rise from 22.70% to 67.13% in English and from 22.11% to 65.68% in Farsi, an increase of over 40 percentage points. GPT is the least influenced, with hint matches staying below 50%, though its accuracy still drops. The English–Farsi gap in hint matches is about 10% for Gemini and smaller for other models. DeepSeek shows the largest shift, with hint matches increasing by +50% in Farsi and +55% in English, despite often labeling hints as False or Uncertain, especially in English. This suggests that even when a model rejects a statement in isolation, it can still be biased by it in a QA context. Claude-3.7 and DeepSeek-V3 also show strong hint sensitivity in Farsi, while GPT-4o and LLaMA-3.1-8B are less prone to follow hints directly but still lose accuracy when hints are added. Although models like LLaMA and DeepSeek differ in how they classify hints and explanations across languages, their overall hint susceptibility is similar, underscoring the impact of added information. Finally, proprietary models show smaller accuracy drops compared to open-source ones, suggesting open-source models are more affected by misleading hints—even when their answers don't align with the hinted option.

Model	**Type**	**Farsi**			**English**		
		T	F	U	T	F	U
Claude	exp	19.9	44.8	50.8	22.0	52.8	46.3
	hint	42.1	44.5	55.6	35.3	48.1	53.6
GPT	exp	44.8	68.0	70.5	31.8	50.6	54.0
	hint	21.9	26.9	25.1	33.2	41.4	43.8
Gemini	exp	18.1	48.0	56.3	26.6	53.4	55.3
	hint	31.4	42.2	56.1	38.6	44.7	50.2
DeepSeek	exp	33.1	56.1	49.7	25.4	40.0	53.4
	hint	49.9	44.6	60.5	45.6	61.7	62.4
Qwen	exp	36.8	59.3	64.0	34.5	60.6	54.1
	hint	34.6	55.6	62.8	32.8	44.2	51.3
Gemma	exp	42.9	54.6	57.3	48.0	73.7	53.0
	hint	42.3	47.2	48.4	42.7	39.1	48.7
Mistral	exp	26.5	-	50.7	17.5	54.8	56.3
	hint	31.9	-	48.1	29.8	23.5	50.6
LLaMA	exp	53.1	53.4	-	55.8	62.7	67.4
	hint	52.6	52.9	-	46.8	40.9	50.0
Command	exp	50.4	66.0	44.8	44.8	53.3	59.7
	hint	43.5	50.9	48.7	34.4	44.4	46.4
Average	exp	36.8	56.5	55.3	34.7	56.3	55.5
	hint	38.7	46.0	50.8	37.6	43.6	50.4

Table 4: Based on the prior tests and without new runs, the table shows the percentages of time, for each of True, False, and Uncertain classes, when including the additional information (either explanation or hint) results in a change of the model's selected option compared to when no additional information is included, divided by the times when that specific label was selected.

Comparing Hints vs. Explanations. Hints often supply partial or tangential details, omitting an explicit statement that any particular option is correct. In Figure 1, for instance, the explanation explicitly states "Josephine was Jamalzadeh's first wife", whereas the hint references his later marriage without clarifying the order of his marriages. These subtle framing choices significantly affect model decisions: hints can misdirect a response, whereas explanations clearly confirm the intended answer.

Additionally, Table 4 shows that models are

more affected by a hint or explanation, when they classify these additional facts as False or Uncertain in isolation. Table 4 does not introduce a new task, but reuses the same QA experiments as in previous tables. For each model and language, we (i) first run QA without any extra information, (ii) then run QA again with either the explanation or the hint added to the prompt, and (iii) check whether the chosen option changes. We then group questions by how the model had classified that statement in the factuality task (True / False / Uncertain), and Table 4 reports, for each group, the percentage of questions where the model's answer changes when the statement is added. In fact, on average, this effect is around 50% for False and Uncertain labels across the languages, while it is around 35% for the True label. These findings emphasize the fact that these additional pieces of information are treated differently depending on the context. We also need to recall that the cases where no change occurred in the model's responses were not included.

Discussion. Our findings highlight two key points. First, LLMs still struggle with factual recognition alone, with True-label rates below 50% in factuality classification. Second, context strongly shapes QA results: accurate explanations can raise accuracy above 90%, while factual but irrelevant hints often reduce performance or bias responses. Although some models vary across languages in factuality classification, their QA results are more consistent. Overall, this shows the impact of domain familiarity, here, Iranian content, and the influence of added context.

Moreover, our intended claim is more specific than "models are always biased by any context". We show that when a truthful-but-misleading hint is the only additional context, it can substantially reduce accuracy and pull models toward the hinted option. Concretely, in Table 3, the average baseline accuracy is around 48% in both languages, while under QA+hint, it drops to about 23–25%, with hint-aligned choices rising from almost 22% to over 65%. This pattern appears across models and categories and shows that factual but irrelevant hints often reduce performance or bias responses.

Moreover, all questions in TRUTHTRAP are answerable by construction: each item is derived from a Farsi Wikipedia page and paired with a manually verified explanation that directly encodes the correct answer. The baseline accuracies and the fact that QA+explanation accuracy exceeds 90% for almost all models and both languages confirm that, once given this explanation, models can almost always find the right answer. The low factuality classification scores (Table 2) are therefore not evidence that questions are unanswerable; rather, they indicate that models often mislabel true statements as False or Uncertain when seen in isolation. This is exactly the contrast we try to highlight: models can benefit from the same facts in QA but struggle to recognize their truth value standalone. Even if the results are expected, we still need to measure the impact of true-but-misleading information. A user who asks a question based on text pulled from a reliable source like Wikipedia won't necessarily know that the information is incomplete. On the surface, these partial snippets look completely relevant, so the model will respond to them, and because they appear related and come from a trusted source, the user may believe the answer without realizing the gaps.

6 Conclusion

We present a bilingual (Farsi–English) dataset to study how LLMs handle factually correct but misleading information in question-answering tasks. As LLMs increasingly address information-seeking questions, they must navigate diverse and sometimes distracting details that can weaken their responses. Our experiments show that LLMs are broadly vulnerable to extra, irrelevant information, regardless of question type or language. Interestingly, they may also label true statements as false when taken out of context, highlighting the importance of contextual framing. Open-source models are more affected by persuasive yet irrelevant content, while proprietary ones are more influenced by explanations. GPT shows relatively stronger resilience, though its accuracy still declines with misleading cues. Overall, our results stress the need to evaluate how LLMs manage accurate but off-target details. Future work should focus on improving context-filtering and verification to boost reliability across languages and domains.

Limitations

Our work is not without limitations, and we acknowledge several constraints that may affect the interpretation and generalizability of our findings. First, the scope of our resource is primarily centered around questions related to Iran. While this focus allowed us to explore the subject matter in

depth, it also means that our conclusions may not readily extend to questions or contexts involving other countries. Future work should consider expanding the geographic and topical diversity of the dataset to improve the applicability of the findings on a global scale. Also, cross-lingual differences should be interpreted with the caveat that English is obtained via post-edited translation from Farsi.

Second, the overall size of our dataset is 1,000 Farsi samples and 1,000 English samples created by translating the Farsi version. This may be considered limited for tasks requiring robust generalization. This constraint is largely due to the challenges involved in generating high-quality, persuasive, and factually accurate information. The process of creating such content is time-consuming and requires careful curation, making it difficult to scale effectively across a wider range of questions.

Finally, we do not explore additional improvement directions. For instance, chain-of-thought prompting could plausibly improve performance, and model adaptation methods (e.g., fine-tuning or weight editing) may further benefit our tasks. We leave these extensions to future work due to scope and resource constraints.

References

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are LLMs to influence in prompts? In *First Conference on Language Modeling*.

Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. Leancontext: Cost-efficient domain-specific question answering using llms. *Natural Language Processing Journal*, 7:100065.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504, Online. Association for Computational Linguistics.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Zahra Bokaei, Walid Magdy, and Bonnie Webber. 2025. Culture matters in toxic language detection in Persian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9290–9304, Vienna, Austria. Association for Computational Linguistics.

Carlos Carrasco-Farre. 2024. Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint arXiv:2404.09329*.

Tyler A Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, and 1 others. 2025. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*.

Gobinda Chowdhury and Sudatta Chowdhury. 2024. Ai-and llm-driven search tools: A paradigm shift in information access for education and research. *Journal of Information Science*, page 01655515241284046.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Pedro Henrique Luz de Araujo and Benjamin Roth. 2024. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *arXiv preprint arXiv:2407.02099*.

Karen de Souza, Alexandre Nikolaev, and Maarit Koponen. 2025. Generative ai for technical writing: Comparing human and llm assessments of generated content. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 661–679.

Saba Emami and Maedeh Mosharraf. 2023. Farcqa: A farsi community dataset for question classification and answer selection. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 567–572.

E Eskola. 1998. University students' information seeking behaviour in a changing learning environment. how are students' information needs, seeking and use affected by new teaching methods. *Information Research*, 4(2):4–2.

Marcos Fernández-Pichel, Juan C Pichel, and David E Losada. 2024. Search engines, llms or both? evaluating information seeking strategies for answering health questions. *arXiv preprint arXiv:2407.12468*.

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? *Preprint*, arXiv:2404.06644.

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. 2024. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1908–1925, Miami, Florida, USA. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jia Hua Jeng, Gloria Anne Babile Kasangu, Alain D Starke, Erik Knudsen, and Christoph Trattner. 2024. Negativity sells? using an llm to affectively reframe news articles in a recommender system. In *ACM Conference on Recommender Systems (RecSys' 24)*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

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 Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Bryan Li, Aleksey Panasyuk, and Chris Callison-Burch. 2024. Uncovering differences in persuasive language in Russian versus English Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, pages 21–35, Miami, Florida, USA. Association for Computational Linguistics.

Louise Limberg and Olof Sundin. 2006. Teaching information seeking: relating information literacy education to theories of information behaviour. *Information Research: an international electronic journal*, 12(1):n1.

Siyi Liu, Qiang Ning, Kishaloy Halder, Zheng Qi, Wei Xiao, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2025. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1838–1854, Albuquerque, New Mexico. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. 2024. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975.

Nishith Reddy Mannuru, Aashrith Mannuru, and Brady Lund. 2024. Large language models (llms) as a tool to facilitate information seeking behavior. *InfoScience Trends*, 1(3):34–42.

AI Meta, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2.

Jyoti Mishra, David Allen, and Alan Pearman. 2015. Information seeking, use, and decision making. *Journal of the association for information science and technology*, 66(4):662–673.

Joao Monteiro, Pierre-Andre Noel, Etienne Marcotte, Sai Rajeswar Mudumba, Valentina Zantedeschi, David Vazquez, Nicolas Chapados, Chris Pal, and Perouz Taslakian. 2024. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. *Advances in Neural Information Processing Systems*, 37:24242–24276.

Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. PerCul: A story-driven cultural evaluation of LLMs in Persian. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.

S Narayanan, William Bailey, Juee Tendulkar, Karen Wilson, Raymond Daley, and Daniel Pliske. 1999. Modeling real-world information seeking in a corporate environment. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 9(2):203–229.

Yishai Ofran, Ora Paltiel, Dan Pelleg, Jacob M Rowe, and Elad Yom-Tov. 2012. Patterns of information-seeking for cancer on the internet: an analysis of real world data.

Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. 2024. Uncovering limitations of large language models in information seeking from tables. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1388–1409, Bangkok, Thailand. Association for Computational Linguistics.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

J Perchik. 2023. Does chatgpt pass the lirads test? comparing quality of ai generated impressions to human reports. *J Gastro Hepato*, 10(5):1–5.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your llm must learn the persian art of taarof. *arXiv preprint arXiv:2509.01035*.

Till Raphael Saenger, Musashi Hinck, Justin Grimmer, and Brandon M. Stewart. 2024. AutoPersuade: A framework for evaluating and explaining persuasive arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16325–16342, Miami, Florida, USA. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, and Debora Nozza. 2025a. Measuring gender bias in language models in Farsi. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 228–241, Vienna, Austria. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025b. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6060–6074, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. *Advances in neural information processing systems*, 36:72044–72057.

Tefko Saracevic, Paul Kantor, Alice Y Chamis, and Donna Trivison. 1988. A study of information seeking and retrieving. i. background and methodology. *Journal of the American Society for Information science*, 39(3):161–176.

Joshua M Scacco and Ashley Muddiman. 2020. The curiosity effect: Information seeking in the contemporary news environment. *New Media & Society*, 22(3):429–448.

Mohammadamin Shafiei, Hamidreza Saffari, and Nafise Sadat Moosavi. 2025. MultiHoax: A dataset of multi-hop false-premise questions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10169–10187, Vienna, Austria. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Lieke LF van Lieshout, Floris P de Lange, and Roshan Cools. 2020. Why so curious? quantifying mechanisms of information seeking. *Current Opinion in Behavioral Sciences*, 35:112–117.

Krithik Vishwanath, Anton Alyakin, Daniel Alexander Alber, Jin Vivian Lee, Douglas Kondziolka, and Eric Karl Oermann. 2025. Medical large language models are easily distracted. *arXiv preprint arXiv:2504.01201*.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Hye Sun Yun and Timothy Bickmore. Online health information seeking in the era of large language models: Cross-sectional online survey study.

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, and Marco Gori. 2025. PersianMCQ-instruct: A comprehensive resource for generating multiple-choice questions in Persian. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 344–372, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Appendix: Dataset Generation Process and Dataset Details

A.1 Dataset Details

This section provides further details regarding the dataset, including the distribution of each sub-category as well as the number of times when the hint targets the correct answer and the number of times it targets a wrong option. The sub-category counts are provided in Table 7 while the target distributions are present in Table 6.

Question class	Count
Date or Time	198
Location	178
Number	165
Person	151
Other proper and common nouns	109
Group or Organization	105
Other	42
Event	27
Artwork	25
Sum	**1000**

Table 5: Distribution of question classes in our final set of 1,000 curated QA items. Each question is assigned a class (e.g., Date/Time, Location, Person), reflecting the principal type of information sought by the query.

Target	Count
About an option other than the answer	956
About the answer	44

Table 6: Distribution of targets of the hints in the dataset. About an option other than the answer means that the hint is talking about one of the wrong options. On the contrary, the answer means that the hint is about the specific correct answer to the query. A hint either (i) targets the correct answer (e.g., negatively phrased questions where the hint talks about the answer but still misleads), or (ii) targets one of the incorrect options by making it look like the correct answer (the majority of cases).

A.2 Dataset Generation

We used a prompt to collect relevant Wikipedia documents for each category. To this end, we simply provided the definitions of each category's three sub-categories and asked the model, GPT4o with search ability, to provide five related documents for each sub-category. This prompt is available in the Table 8. To collect the documents, we used the following category and sub-category definitions:

Food

- **Cuisine**: Signature dishes, cooking styles, traditional meals.
- **Ingredients**: Locally grown spices, crops, and special ingredients.

Category	Subcategory	Count
Arts and Literature	Artists	28
	Books	18
	Writers	54
Education	Education System and Literacy	51
	Famous Educators	34
	Schools and Universities and Curriculum	15
Entertainment	Cinema and TV	32
	Music	37
	Others	31
Food	Cuisine	44
	Drinks	34
	Ingredients	22
Geography	Cities and Regions	24
	Geopolitics	19
	Natural Features and Resources	57
History	Historical Figures	56
	Important Events	22
	Landmarks	22
Holidays, Celebrations, Leisure	Festivals	57
	National Holidays	21
	Others	22
Religion	Holy Sites	34
	Others	17
	Religions and Religious Practices	49
Science and Technology	Engineering	17
	Others	39
	Scientists	44
Sports	Athletes	28
	National and Popular Sports	48
	Tournaments and Sports Venues	24

Table 7: Distribution of counts across categories and subcategories

- **Drinks**: Popular beverages, traditional teas, or alcoholic drinks.

Sports

- **National and Popular Sports**: Widely played or watched sports in the country and Official sports of a country.
- **Athletes**: Famous sportspeople or Olympic medalists.
- **Tournaments and Sports Venues**: Major leagues, championships, or cups, as well as iconic stadiums, arenas, or tracks.

Education

- **Education System AND Literacy**: Structure (primary, secondary, higher education) AND Efforts to promote literacy or improve access to education.
- **Schools and Universities AND Curriculum**: Prestigious or historic institutions AND Subjects emphasized or unique courses.
- **Famous Educators**: Scholars, reformers, or pioneers in education.

Holidays/Celebrations/Leisure

- **National Holidays**: Independence days, constitution days, or memorials.
- **Festivals**: Cultural, religious, or seasonal festivals.
- **Others**: Other topics related to Holidays/Celebrations/Leisure.

History

- **Historical Figures**: Leaders, revolutionaries, empires and kingdoms, or intellectuals.
- **Important Events**: Battles, treaties, or turning points in history.
- **Landmarks**: Historical monuments or UNESCO heritage sites.

Geography

- **Natural Features AND Resources**: Mountains, rivers, lakes, and deserts AND Natural resources, agriculture, or energy production.
- **Cities AND Regions**: Capitals, major cities, or urban landmarks AND Administrative divisions or cultural regions.
- **Geopolitics**: Borders, neighbors, or disputed territories.

Science and Technology

- **Scientists**: Modern renowned scientists.
- **Engineering**: Famous modern constructions, bridges, or technology.
- **Others**: Other related topics to Science and Technology like Medical Breakthroughs, Research Centers, Computing Pioneers, Green Technology, Digital Platforms, and Communications.

Arts and Literature

- **Writers**: Prominent authors, poets, or playwrights.
- **Books**: National epics, famous novels, or historical documents.
- **Artists**: Prominent artists.

Religion

- **Religions and Religious Practices**: Popular religions and Worship styles or Religious rituals.
- **Holy Sites**: Temples, churches, mosques, or pilgrimage locations.
- **Others**: Religious Leaders, Religious Festivals, or Sacred Texts.

Entertainment

- **Cinema and TV**: National cinema, famous directors, popular movies, or actors.
- **Music**: Traditional music styles, Musicians, or iconic bands.
- **Others**: Other topics related to entertainment like theater, gaming, festivals related to entertainment, or media.

After the collection step, we prompted Claude sonnet 3.7, which has been proven to be a strong model in Farsi based on the previous work such as (Moosavi Monazzah et al., 2025) and (Saffari et al., 2025b). Accordingly, Table 9 shows the prompt that we used to generate the very initial version of our dataset, with which the content of each document was given. Moreover, we used the English language for some parts of the prompts, where we are explaining the ISQ concept as well as the output structure. This is due to the fact that models show better instruction-following abilities in English in such cases. However, since the provided content and also the generated question were supposed to be in Farsi, we gave examples in Farsi. We tested some other combinations of these two languages in the generation prompt, from relying solely on Farsi to using English, and we got the best results with the current bilingual one.

A.3 Annotation Guidelines And Statistics

As explained in the main text, two annotators went through the resource, one by one. The annotations process resulted in 1,000 samples. There was no conflict concerning the annotations of the two annotators, except for the question class.

Once the first annotator completed this multi-step protocol and finalized the 1,000 items, the second annotator performed an independent review, confirming the correctness of explanations, hints, and distractors. Across all items, only 23 disagreements arose concerning question-type labels, often when one annotator used a more specific category (e.g., date/time) while the other used a broader one (e.g., proper/common noun). These discrepancies were resolved via brief discussion, generally favoring the finer-grained classification. No other issues emerged.

B Appendix: Additional Results

In this section, we present additional results beyond what we have already presented in the main paper. Tables 11-19 show the details of the results for each category.

Moreover, in Tables 11–20, answer_match is, as in Table 3, the percentage of questions in that category where the model selects the correct option. The Hint_match(Target=1) columns report, for each QA setup (Baseline / Explanation / Hint), the percentage of questions in that category whose hint targets an incorrect option and

I am collecting Wikipedia documents related to Iran and different categories. Currently, I am focusing on category with three sub-categories.

- sub-category 1: definition
- sub-category 2: definition
- sub-category 3: definition

Now, give me 5 important Wikipedia documents for each sub-category related to Iran. Provide the links to these pages.
THE PAGES MUST BE IN THE PERSIAN LANGUAGE

Table 8: The prompt used for collecting Wikipedia documents.

where the model chooses that hinted option. The Hint_match(Target=0) columns use a different denominator: they consider only questions whose hint describes the correct answer in a less appealing way and measure how often the model chooses a wrong option. Because Target=0 and Target=1 are disjoint subsets with different sizes, the values under answer_match, Hint_match(Target=0), and Hint_match(Target=1) are not meant to sum to 100 and can exceed 100 if added. The outlier values (e.g., columns that are mostly 0.00 but 100 for the Holiday category) arise from very small sample sizes in some category–target combinations, so a few successes or failures produce extreme percentages. Our main quantitative claims rely on the aggregated Table 3 and Table 4, while Tables 11–20 are intended as additional diagnostic detail.

We are developing a dataset of information-seeking questions for Iran. We also want to include a persuasive hint per question about one of the other false options, making it more attractive compared to the true answer. This persuasive hint is TRUE, but it makes the wrong option more appealing.

Here are some example questions with persuasive hints:

- کدام استان ایران پرجمعیت تر است؛ آذربایجان غربی یا کرمان؟
- آذربایجان غربی | کرمان
- **راهنمایی**: کرمان بزرگترین استان ایران است.
- **توضیح**: با وجود اینکه کرمان استان بزرگتری به لحاظ مساحت است؛ این استان جمعیت کمتری از آذربایجان غربی دارد.

- کدام ماده اصلی در تهیه میرزا قاسمی گیلان استفاده نمی‌شود؟
- بادمجان | تخم مرغ | سیر | نعنا
- **راهنمایی**: از آنجایی که سیر باعث بد بو شدن دهان می شود در غذاهای ایرانی کمتر استفاده می شود.
- **توضیح**: با وحود اینکه سیر میتواند باعث بد بو شدن دهان می شود؛ سیر از مواد اصلی میرزا قاسمی است.

- کدام رودخانه ایرانی طولانی تر است, سفید رود یا اترک؟
- سفید رود | اترک
- **راهنمایی**: اترک با 563 کیلومتر طول از طولانی ترین رودخانه های ایران است.
- **توضیح**: اترک و سفید رود هر دو از رودخانه های طولانی ایران هستند. با این وجود؛ سفید رود با طول 670 کیلومتر از اترک 563 کیلومتری طولانی تر است.

- کدام آش مخصوص دوران سرماخوردگی است؟
- آش انار | آش دنگو | آش کدو | آش ماست
- **راهنمایی**: آش انار به دلیل داشتن ویتامین C فراوان برای سیستم ایمینی بدن بسیار مفید است.
- **توضیح**: با وجود اینکه آش انار میتواند برای سرماخوردگی مفید باشد؛ آش دنگو در فصل زمستان برای مقابله با سرماخوردگی تهیه می‌شود.

- ریشه واژه "آش" از کدام زبان است؟
- عربی | ترکی | پیشاهندواروپایی | پهلوی
- **راهنمایی**: زبان پهلوی منشا بسیاری از واژه های مربوط به غذا در ایران است.
- **توضیح**: واژهٔ «آش» در زبان پارسی از ریشهٔ پیشاهندواروپایی «*ed* /hed*» به معنای «خوردن» مشتق شده است.

Remember that our focus is on FACTUAL questions.
Remember that hints are TRUE.

Return each question in the following format:
<question> | <first_option> | <second_option> | <third_option> | <fourth_option> | <answer_index> | <description> | <hint>

Table 9: The prompt for generating the initial resource. The Farsi parts of the prompt present three examples of the type of question that are requested to generate.

ساده:

سوال: [سوال]

1: [گزینه اول]

2: [گزینه دوم]

3: [گزینه سوم]

4: [گزینه چهارم]

فقط شماره ی گزینه درست را بدون هیچ توضیح اضافه ای برگردان.

سوال به همراه توضیح:

سوال: [سوال]

1: [گزینه اول]

2: [گزینه دوم]

3: [گزینه سوم]

4: [گزینه چهارم]

:یک قطعه اطلاعات

[توضیح]

فقط شماره ی گزینه درست را بدون هیچ توضیح اضافه ای برگردان.

سوال به همراه راهنمایی:

سوال: [سوال]

1: [گزینه اول]

2: [گزینه دوم]

3: [گزینه سوم]

4: [گزینه چهارم]

:یک قطعه اطلاعات

[راهنما]

فقط شماره ی گزینه درست را بدون هیچ توضیح اضافه ای برگردان.

تشخیص درستیِ راهنما:

آیا این جمله از نظر واقعیت درست است، نادرست است، یا نامشخص و به طور قطعی نمیتوان نظر داد؟
"[راهنما]" | فقط یک کلمه از بین ['درست', 'نادرست', 'نامشخص'] را بدون هیچ توضیح اضافه ای برگردان

تشخیص درستی توضیح:

آیا این جمله از نظر واقعیت درست است، نادرست است، یا نامشخص و به طور قطعی نمیتوان نظر داد؟
"[توضیح]" | فقط یک کلمه از بین ['درست', 'نادرست', 'نامشخص'] را بدون هیچ توضیح اضافه ای برگردان

Table 10: The Farsi counterparts of the prompts provided in the table 1 for experiments.

Table 11: Results for GPT. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. A hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	60.20	45.00	100	100	45.00	45.00	33.33	66.67	0.00	0.00	66.67	66.67	15.79	38.14	0.00	0.00	38.14	42.27
Educ.	51.52	44.00	96.00	98.00	44.00	39.00	66.67	33.33	0.00	0.00	33.33	66.67	18.75	42.27	2.06	2.06	42.27	49.48
Ent	55.56	40.00	100	99.00	40.00	37.00	100	100	0.00	0.00	100	100	18.37	42.42	0.00	0.00	42.42	50.51
Food	54.55	38.00	96.00	97.00	38.00	32.00	33.33	33.33	0.00	0.00	33.33	40.00	19.05	47.06	3.53	2.35	47.06	50.59
Geo	50.00	45.00	99.00	98.00	45.00	43.43	0.00	16.67	0.00	0.00	16.67	33.33	11.70	39.36	1.06	1.06	39.36	44.09
Hist	70.41	52.00	99.00	99.00	52.00	51.00	0.00	20.00	0.00	0.00	20.00	20.00	12.90	33.68	1.05	1.05	33.68	43.16
Holiday.	47.42	31.00	95.00	96.00	31.00	27.55	100	100	100	66.67	100	66.67	22.34	57.73	0.00	0.00	57.73	57.89
Rel.	67.35	53.00	99.00	99.00	53.00	50.00	25.00	25.00	0.00	0.00	25.00	25.00	12.77	37.50	1.04	1.04	37.50	40.62
Sports	55.10	44.00	100	99.00	44.00	37.00	0.00	0.00	0.00	0.00	0.00	0.00	22.45	41.00	0.00	0.00	41.00	55.00
Sci_Tech	62.00	47.96	98.00	99.00	47.96	46.00	25.00	25.00	0.00	0.00	25.00	25.00	18.75	39.36	2.08	1.04	39.36	42.71
Overall	57.40	43.99	98.20	98.40	43.99	40.82	31.82	36.36	6.82	4.55	36.36	40.91	17.30	41.82	1.05	0.84	41.82	47.64

Table 12: Results for Claude Sonnet. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	60.00	71.00	100	100	27.00	32.00	66.67	33.33	0.00	0.00	66.67	66.67	18.56	12.37	0.00	0.00	61.86	59.79
Educ.	52.58	60.20	100	99.00	18.56	28.00	0.00	33.33	0.00	0.00	0.00	0.00	23.40	17.89	0.00	1.03	77.66	69.07
Ent	58.51	58.16	99.00	99.00	20.62	21.65	100	100	100	0.00	0.00	100	17.20	14.43	0.00	0.00	77.08	72.92
Food	52.00	60.20	97.00	98.00	25.51	28.87	40.00	26.67	0.00	0.00	26.67	33.33	21.18	13.25	3.53	2.35	75.90	67.07
Geo	61.00	62.00	99.00	99.00	23.23	31.00	0.00	16.67	0.00	0.00	0.00	16.67	17.02	17.02	1.06	1.06	67.74	61.70
Hist	58.00	67.00	99.00	99.00	32.32	43.00	20.00	20.00	0.00	0.00	20.00	20.00	22.11	15.79	1.05	1.05	62.77	53.68
Holiday.	55.00	57.00	95.00	97.00	19.00	27.00	100	66.67	100	66.67	66.67	33.33	15.46	25.77	1.03	0.00	74.23	69.07
Rel.	66.67	68.00	99.00	100	33.67	41.00	25.00	50.00	0.00	0.00	25.00	50.00	15.79	10.42	1.04	0.00	62.77	54.17
Sports	56.57	57.00	100	100	18.18	21.00	0.00	0.00	0.00	0.00	0.00	0.00	22.22	20.00	0.00	0.00	68.69	75.00
Sci_Tech	62.50	60.82	98.00	100	27.27	21.88	25.00	25.00	25.00	0.00	0.00	25.00	18.48	18.28	1.04	0.00	69.47	77.17
Overall	58.27	62.16	98.60	99.10	24.54	29.60	34.09	31.82	11.36	4.55	22.73	31.82	19.13	16.58	0.84	0.52	69.75	65.96

Table 13: Results for CommandR. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	37.00	26.00	96.00	87.00	11.00	12.00	33.30	33.30	0.00	33.30	0.00	0.00	21.60	26.80	0.00	6.20	75.30	80.40
Educ.	40.00	42.00	99.00	90.00	13.00	15.00	66.70	100	0.00	66.70	66.70	66.70	28.90	25.80	0.00	2.10	73.20	67.00
Ent	30.00	36.00	96.00	89.00	8.00	20.00	100	100	0.00	0.00	100	0.00	32.30	18.20	2.00	2.00	81.80	59.60
Food	34.00	40.00	87.00	79.80	23.00	21.00	73.30	53.30	46.70	46.70	46.70	40.00	24.70	17.60	3.50	6.00	61.20	65.90
Geo	33.00	43.00	94.00	84.00	13.00	15.00	66.70	50.00	16.70	33.30	66.70	33.30	37.20	24.50	3.20	6.40	74.50	74.50
Hist	40.00	30.00	94.00	84.00	22.20	17.00	60.00	80.00	0.00	20.00	80.00	40.00	24.20	23.20	4.20	7.40	64.90	68.40
Holiday.	29.00	26.00	86.00	82.00	12.00	14.00	66.70	100	100	66.70	33.30	66.70	25.80	27.80	3.10	5.20	73.20	69.10
Rel.	49.00	36.00	94.00	91.00	25.00	22.00	75.00	100	25.00	50.00	25.00	0.00	26.00	30.20	4.20	3.10	65.60	67.70
Sports	38.00	35.00	95.00	93.00	10.00	15.00	0.00	0.00	0.00	0.00	0.00	0.00	27.00	23.00	1.00	3.00	79.00	74.00
Sci_Tech	35.00	43.00	95.00	87.00	19.00	15.00	50.00	25.00	50.00	50.00	50.00	25.00	30.20	19.80	2.10	7.30	74.00	74.00
Overall	36.50	35.70	93.60	86.70	15.60	16.60	65.90	63.60	31.80	43.20	50.00	34.10	27.80	23.70	2.30	4.80	72.50	70.10

Table 14: Results for Deepseek. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	56.00	66.00	99.00	100	19.00	28.00	33.33	33.33	0.00	0.00	0.00	33.33	16.49	13.40	0.00	0.00	77.32	63.92
Educ.	53.00	56.00	100	99.00	17.17	21.00	66.67	100	0.00	0.00	33.33	66.67	18.56	21.65	0.00	1.03	80.21	72.16
Ent	51.00	53.00	100	99.00	8.00	14.00	100	100	0.00	0.00	0.00	100	21.21	18.18	0.00	0.00	91.92	81.82
Food	52.00	53.00	95.00	97.00	19.00	23.00	53.33	40.00	13.33	0.00	26.67	26.67	23.53	21.18	3.53	3.53	81.18	80.00
Geo	57.00	58.00	99.00	99.00	17.17	20.00	33.33	16.67	0.00	0.00	0.00	0.00	17.02	10.64	1.06	1.06	77.42	72.34
Hist	57.00	61.00	98.00	97.00	25.00	38.00	60.00	40.00	20.00	20.00	40.00	0.00	18.95	13.68	1.05	2.11	67.37	61.05
Holiday.	42.00	49.00	96.00	96.00	11.00	21.00	100	66.67	100	66.67	33.33	66.67	23.71	18.56	1.03	1.03	84.54	74.23
Rel.	59.00	68.00	99.00	99.00	25.00	35.00	25.00	25.00	0.00	0.00	25.00	0.00	15.62	15.62	1.04	1.04	73.96	61.46
Sports	47.00	58.00	100	98.00	10.00	15.00	0.00	0.00	0.00	0.00	0.00	0.00	26.00	17.00	0.00	0.00	82.00	78.00
Sci_Tech	56.00	57.00	99.00	99.00	19.00	20.00	50.00	25.00	0.00	0.00	0.00	0.00	15.62	21.88	0.00	1.04	83.33	77.08
Overall	53.00	57.90	98.50	98.30	17.03	23.50	52.27	40.91	13.64	6.82	20.45	22.73	19.67	17.15	0.73	1.05	79.98	72.18

Table 15: Results for Gemini. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	59.60	70.00	100	100	32.00	38.40	0.00	0.00	0.00	0.00	33.30	33.30	20.80	11.30	0.00	0.00	58.50	54.20
Educ.	51.00	61.00	99.00	100	22.40	34.00	33.30	66.70	0.00	0.00	0.00	33.30	23.70	16.50	0.00	0.00	70.50	62.80
Ent	51.00	67.70	99.00	99.00	24.50	40.40	100	100	100	0.00	100	100	23.20	11.20	0.00	0.00	68.00	54.10
Food	55.00	52.00	95.00	96.00	26.50	27.30	33.30	26.70	6.70	6.70	26.70	33.30	14.10	15.30	3.50	2.40	71.10	63.10
Geo	54.00	61.20	99.00	99.00	19.80	28.30	16.70	33.30	0.00	0.00	0.00	16.70	18.10	14.10	1.10	1.10	73.30	59.10
Hist	56.60	65.30	98.00	97.00	32.70	45.00	20.00	40.00	0.00	20.00	20.00	40.00	19.10	16.10	2.10	2.10	57.00	41.10
Holiday.	47.00	46.90	93.00	98.00	22.40	28.00	100	66.70	100	66.70	66.70	33.30	23.70	21.10	2.10	0.00	69.50	66.00
Rel.	65.00	63.00	100	100	37.40	43.00	25.00	25.00	0.00	0.00	25.00	25.00	17.70	16.70	0.00	0.00	57.90	47.90
Sports	46.00	59.20	100	100	23.20	35.00	0.00	0.00	0.00	0.00	0.00	0.00	26.00	16.30	0.00	0.00	64.60	49.00
Sci_Tech	55.00	67.30	99.00	99.00	22.30	25.50	50.00	0.00	0.00	0.00	0.00	0.00	13.50	18.10	0.00	1.00	72.20	72.30
Overall	54.00	61.40	98.20	98.80	26.40	34.50	34.10	31.80	11.40	9.10	22.70	29.50	20.10	15.70	0.80	0.60	66.20	56.80

Table 16: Results for Gemma. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	46.00	46.50	100	99.00	15.00	19.00	66.70	66.70	0.00	0.00	33.30	66.70	15.50	19.80	0.00	0.00	70.10	69.10
Educ.	44.00	44.00	99.00	97.00	14.00	16.00	33.30	66.70	0.00	0.00	33.30	66.70	26.80	26.80	1.00	2.10	74.20	69.10
Ent	39.00	53.00	98.00	98.00	14.10	22.00	100	100	0.00	100	100	100	30.30	22.20	0.00	0.00	70.40	69.70
Food	48.00	53.00	95.90	96.00	22.40	23.20	40.00	26.70	6.70	6.70	60.00	53.30	17.60	20.00	3.60	3.60	67.50	64.30
Geo	45.00	52.00	97.00	97.00	21.00	22.00	33.30	33.30	16.70	16.70	50.00	50.00	18.10	12.80	1.10	1.10	58.50	61.70
Hist	44.00	50.00	96.00	97.00	21.00	23.20	80.00	60.00	40.00	20.00	80.00	60.00	24.20	20.00	2.10	2.10	63.20	67.00
Holiday.	40.00	47.00	95.00	97.00	15.00	20.00	100	66.70	100	66.70	66.70	33.30	22.70	16.50	2.10	1.00	71.10	70.10
Rel.	56.00	60.00	99.00	98.00	32.30	39.00	25.00	25.00	0.00	0.00	25.00	50.00	25.00	17.70	1.00	1.00	62.10	53.10
Sports	50.50	53.00	98.00	99.00	21.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	19.20	24.00	0.00	0.00	67.00	68.00
Sci_Tech	48.50	42.40	98.00	98.00	18.20	20.20	50.00	50.00	0.00	0.00	75.00	50.00	24.20	26.30	1.00	1.00	69.50	70.50
Overall	46.10	50.10	97.60	97.60	19.40	22.50	50.00	43.20	15.90	13.60	56.80	54.50	22.40	20.60	1.20	1.20	67.40	66.30

Table 17: Results for LLaMa. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	30.60	39.00	94.00	94.00	9.00	12.00	100	100	33.30	100	66.70	33.30	26.30	17.50	3.10	0.00	73.20	81.40
Educ.	38.40	43.00	94.00	94.00	15.00	8.00	66.70	66.70	33.30	33.30	66.70	66.70	28.10	32.00	2.10	2.10	79.40	85.60
Ent	40.00	44.00	93.00	97.00	12.00	18.00	0.00	0.00	0.00	0.00	0.00	0.00	19.10	23.20	5.10	2.00	74.70	75.80
Food	48.00	48.00	96.00	86.00	24.00	16.00	46.70	53.30	6.70	53.30	60.00	53.30	21.70	20.00	2.40	0.00	62.40	69.40
Geo	32.30	40.00	93.00	93.00	13.00	15.00	50.00	66.70	16.70	16.70	50.00	33.30	23.70	20.20	3.20	4.30	73.40	74.50
Hist	34.00	40.00	87.00	95.00	17.00	12.00	80.00	40.00	40.00	20.00	20.00	20.00	20.00	17.90	5.30	1.10	65.30	74.70
Holiday.	36.00	38.00	92.00	91.00	12.00	15.00	100	100	100	100	33.30	33.30	18.60	19.60	0.00	2.10	76.30	72.20
Rel.	36.40	41.00	95.00	93.00	22.00	14.00	75.00	100	25.00	100	75.00	75.00	28.40	28.10	0.00	0.00	67.70	72.90
Sports	28.60	36.00	94.00	98.00	12.00	11.00	0.00	0.00	0.00	0.00	0.00	0.00	31.60	27.00	1.00	1.00	70.00	76.00
Sci_Tech	36.80	43.00	89.00	92.00	19.00	13.00	25.00	25.00	50.00	25.00	25.00	50.00	25.30	19.80	5.20	3.10	69.80	79.20
Overall	36.10	41.20	92.70	93.30	15.50	13.40	59.10	61.40	27.30	50.00	50.00	45.50	24.30	22.60	2.70	1.60	71.30	76.30

Table 18: Results for Mistral. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	45.00	34.00	96.00	97.00	17.20	10.00	66.70	66.70	33.30	0.00	66.70	33.30	24.70	19.60	1.00	0.00	67.70	74.20
Educ.	49.00	48.00	98.00	97.00	18.00	19.00	66.70	100	0.00	33.30	66.70	66.70	22.10	24.70	1.00	0.00	73.20	68.00
Ent	47.50	37.00	100	99.00	14.10	14.00	0.00	0.00	0.00	100	100	100	22.40	22.20	0.00	0.00	73.50	76.80
Food	45.50	45.00	96.00	92.00	23.20	24.00	66.70	73.30	6.70	40.00	40.00	33.30	16.70	15.30	3.50	0.00	73.80	71.80
Geo	40.00	39.00	97.00	94.00	13.10	23.00	33.30	66.70	16.70	16.70	16.70	50.00	20.20	19.10	1.10	1.10	76.30	68.10
Hist	41.00	44.00	96.00	93.00	16.20	17.00	60.00	60.00	20.00	20.00	60.00	60.00	21.10	20.00	3.20	4.20	62.80	73.70
Holiday.	46.00	44.00	96.00	92.00	18.00	14.00	100	66.70	100	66.70	66.70	66.70	21.60	24.70	1.00	2.10	73.20	74.20
Rel.	50.00	50.00	98.00	97.00	33.30	37.00	75.00	75.00	0.00	25.00	25.00	25.00	21.90	13.50	1.00	1.00	58.90	55.20
Sports	41.40	45.00	99.00	95.00	17.20	13.00	0.00	0.00	0.00	0.00	0.00	0.00	20.20	19.00	0.00	0.00	69.70	75.00
Sci_Tech	41.20	44.00	97.00	94.00	17.50	14.00	50.00	25.00	0.00	25.00	0.00	25.00	23.70	16.70	1.00	2.10	69.90	78.10
Overall	44.70	43.00	97.30	95.00	18.80	18.50	61.40	65.90	15.90	31.80	40.90	43.20	21.50	19.60	1.30	1.00	69.90	71.50

Table 19: Results for Qwen. In this table, we present the results for different categories, organized by the three QA settings. The answer match means the percentage of times when the model's predicted answer matches the correct option. Hint match generally means that the hint made the model make a mistake. When the target is 1, it means that the hint was about a wrong option, and the model chose that. On the other hand, target 0 means that the hint was about the correct answer, and the model chose something other than that correct answer.

Category	**answer_match**						**Hint_match(Target=0)**						**Hint_match(Target=1)**					
	Bas		Exp		Hin		Bas		Exp		Hin		Bas		Exp		Hin	
	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa	En	Fa
Arts_Lit	39.00	37.00	99.00	99.00	21.00	10.00	66.70	66.70	0.00	0.00	66.70	66.70	24.70	17.50	0.00	0.00	64.90	82.50
Educ.	48.50	44.00	99.00	98.00	24.00	20.00	66.70	66.70	0.00	0.00	33.30	33.30	15.60	20.60	0.00	1.00	67.00	76.30
Ent	46.50	41.00	99.00	98.00	11.00	14.00	100	100	0.00	0.00	100	100	24.50	15.20	0.00	0.00	74.70	77.80
Food	46.50	44.00	96.00	96.00	23.00	23.00	33.30	46.70	6.70	6.70	33.30	26.70	31.00	18.80	3.50	3.50	76.50	75.30
Geo	53.00	48.00	99.00	97.00	25.00	21.00	16.70	50.00	0.00	33.30	16.70	33.30	13.80	20.20	1.10	1.10	66.00	70.20
Hist	47.00	40.00	96.00	97.00	25.00	22.20	60.00	60.00	20.00	20.00	60.00	60.00	20.00	28.40	3.20	2.10	57.90	66.00
Holiday.	41.00	40.00	96.00	97.00	20.00	19.00	100	66.70	100	66.70	66.70	33.30	20.60	21.60	1.00	0.00	71.10	73.20
Rel.	62.00	55.00	99.00	97.00	32.00	29.00	25.00	25.00	0.00	25.00	25.00	25.00	16.70	17.70	1.00	1.00	60.40	63.50
Sports	40.00	34.00	98.00	99.00	13.00	14.10	0.00	0.00	0.00	0.00	0.00	0.00	26.00	27.00	0.00	0.00	72.00	74.70
Sci_Tech	45.90	41.80	98.00	98.00	24.20	14.00	50.00	50.00	0.00	0.00	25.00	25.00	20.20	19.10	1.00	2.10	63.20	80.20
Overall	46.90	42.50	97.90	97.60	21.80	18.60	45.50	52.30	11.40	15.90	38.60	36.40	21.20	20.60	1.00	1.00	67.30	74.00

Model	Lang.	answer_match	hint_matches(Target=0)	hint_matches(Target=1)
Claude	Farsi	99.12	3.86	0.65
	English	98.5	5.20	0.70
GPT	Farsi	98.50	4.55	0.73
	English	98.49	6.82	0.74
Gemini	Farsi	96.80	1.90	9.10
	English	94.60	3.50	11.40
DeepSeek	Farsi	86.30	7.50	25.00
	English	81.8	10.40	22.70
Qwen3	Farsi	83.00	10.20	15.90
	English	74.70	13.30	29.50
Gemma	Farsi	84.50	8.20	29.50
	English	90.20	5.60	29.50
Mistral	Farsi	76.50	11.90	44.20
	English	77.20	11.20	18.20
LLaMA	Farsi	61.80	18.10	47.70
	English	57.30	19.90	38.60
Command	Farsi	60.40	22.00	34.10
	English	69.30	18.90	34.10

Table 20: Model performances when both explanation and hints were included. answer match shows the correct cases, while hint matches show the cases where hints were effective, similar to table 19.

Model	Type	Farsi			English		
		T	F	U	T	F	U
Claude	exp	98.73	98.85	99.31	97.66	98.11	98.91
	hint	57.87	54.88	70.91	54.55	65.38	70.54
GPT	exp	96.67	98.60	98.92	98.58	96.84	98.79
	hint	42.11	42.71	54.39	42.26	35.03	48.86
Gemini	exp	99.20	98.44	98.85	97.40	98.87	98.33
	hint	45.75	55.34	70.73	56.80	62.94	68.77
DeepSeek	exp	98.60	97.56	97.98	96.77	100	99.00
	hint	67.87	57.14	74.02	65.92	80.85	79.43
Qwen	exp	96.36	97.58	97.88	95.69	97.79	98.40
	hint	58.58	72.63	79.49	54.58	60.64	70.44
Gemma	exp	97.31	96.71	98.02	97.23	96.05	97.41
	hint	62.73	61.42	64.71	63.31	54.69	68.54
Mistral	exp	75.51	-	86.75	95.24	96.77	97.24
	hint	50.00	-	61.85	48.25	44.12	69.25
LLaMA	exp	92.09	94.43	-	93.10	90.57	89.80
	hint	73.78	71.49	-	69.12	64.02	64.52
Command	exp	86.19	88.00	85.42	93.01	100	93.59
	hint	67.09	64.91	66.00	67.22	66.67	69.73

Table 21: The table shows the percentages of time, for each of True (T), False (F), and Uncertain (U) classes, when including the additional information (either explanation (exp) or hint) results in a change of the model's selected option compared to when no additional information is included or when there is no change but the baseline answer aligns with the targeted option, divided by the times when that specific label was selected. This could be seen as another version of table 4