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

Anonymous ACL submission

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 persuasive-yet-misleading true hint. We then evaluate five 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 76.0 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://anonymous.4open.science/r/TRUTHTRAP-FC34.

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

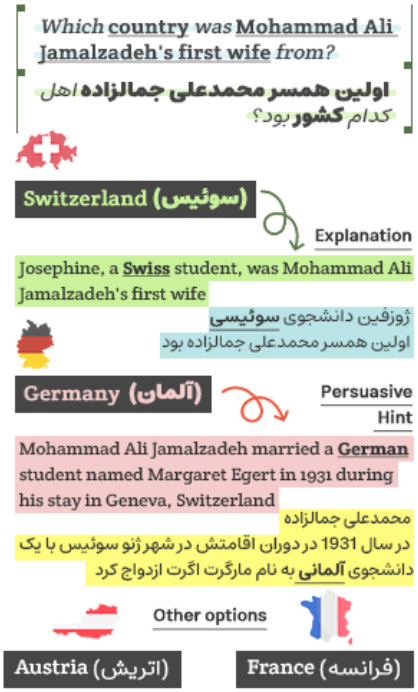

Figure 1: A bilingual QA example from the "Art and Literature" category, showing both English and Farsi question texts ("Which country was Mohammad Ali Jamalzadeh's first wife from?") and four candidate answers (Switzerland, Germany, Austria, France). The correct answer, Switzerland, is supplied by the 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.

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 misleading 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 exceeding 10%. Our error analysis shows that the worst scenario arises when a hint partially overlaps with the correct explanation, causing models to treat both as true but ultimately select the off-target detail.

This paper provides the following contributions: (1) A novel bilingual dataset specifically designed to study the impact of true-but-persuasive hints 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). 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., 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 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, 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, Ge-

ography, 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.

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 generated 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 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.

Models. We evaluated five multilingual LLMs, covering both proprietary and open-source options: Claude Sonnet 3.7 (**Claude-3.7**),[2] **GPT-4o** (Hurst et al., 2024), Qwen2.5-7B-Instruct (**Qwen2.5-7B**) (Yang et al., 2024), LLaMA-3.1-8B-Instruct (**LLaMA-3.1-8B**) (Meta et al., 2024)[3], and **DeepSeek-V3** (Bi et al., 2024). All experiments were conducted in April 2025, 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.

5 Results and Discussion

We present our findings in two parts: factuality classification (Section 5.1), which evaluates how reliably models label individual factual explanation and hint statements as True, False, or Uncer-

[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** and **QA + Hint**, highlighting how each independently influences model performance.

[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 pertaining (Hosseinbeigi et al., 2025; Saffari et al., 2025; Moosavi Monazzah et al., 2025; Zeinalipour et al., 2025).

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

Table 1: English prompt templates for the five experimental modes; Farsi templates follow the same structure.

Model	Type	Farsi			English		
		T	F	U	T	F	U
Qwen2.5-7B	exp	32.8	64.8	2.4	0.1	99.6	0.3
	hint	42.7	51.9	5.4	0	99.9	0.1
GPT-4o	exp	57.2	15.0	27.8	49.4	25.3	25.3
	hint	59.0	17.1	23.9	53.1	23.9	23
DeepSeek-V3	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
Claude-3.7	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
LLaMA-3.1-8B	exp	32.1	66.2	1.7	91.9	7.2	0.9
	hint	36.8	61.0	2.2	93.8	4.7	1.5

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 for each statement type, showing how consistently the systems identify correct statements as True.

tain, 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

Tables 2 summarize how each model labels explanations and hints (both factually correct) in Farsi and English; any choice other than True constitutes a classification error. In Farsi, GPT-4o and DeepSeek-V3 consistently outperform other models, with DeepSeek-V3 leading at 66.3% True labels for hints and GPT-4o nearly matching DeepSeek-V3 on correctly labeling explanations (57.2% vs. 57.1%). By contrast, LLaMA-3.1-8B lags behind in Farsi yet excels in English, surpassing even strong proprietary systems, reflecting its extensive post-training on factual data. Qwen2.5-7B remains the weakest in English, misclassifying most explanations and hints as False. These results highlight a language-based divergence. Some models (e.g., GPT-4o, DeepSeek-V3) achieve higher factual recognition in Farsi, possibly due to extensive pretraining on Farsi data, while LLaMA-3.1-8B shows superior performance in English. 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. Table 3 presents model performance on QA tasks with no additional context. Overall accuracy remains under 50%, averaging 48.1% in English and 47.5% in Farsi, underscoring the dataset's inherent difficulty. GPT-4o, DeepSeek-V3, and Claude-3.7 achieve the highest scores, though 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.

Category	Qwen2.5-7B				GPT-4o				DeepSeek-V3				Claude-3.7				LLaMA-3.1-8B			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Literature	36	26	31	20	59	16	45	37	56	17	66	14	60	20	71	13	41	23	22	28
Education	35	31	36	23	51	20	44	42	53	20	56	24	51	22	59	18	36	28	28	29
Entertainment	32	25	34	24	55	19	40	43	51	21	53	19	55	17	57	15	31	28	27	29
Food	45	23	38	20	54	21	38	45	52	20	53	24	52	18	59	15	35	17	26	34
Geography	37	28	33	28	50	11	45	38	57	16	58	11	61	16	62	17	32	26	23	29
History	37	27	44	21	69	12	52	33	57	18	61	15	58	21	67	15	26	23	30	28
Holidays & Leisure	31	30	34	25	46	21	31	59	42	23	49	20	55	15	57	25	32	27	29	21
Religion	41	34	40	26	66	12	53	37	59	15	68	16	66	15	68	10	42	28	38	28
Science & Technology	43	22	30	22	62	19	47	38	56	15	57	17	60	17	59	17	36	30	26	24
Sports	31	27	40	20	54	22	44	41	47	26	58	17	56	22	57	20	41	21	36	20
Average Accuracy (%)	36.8	27.3	37.0	22.9	56.6	18.3	43.9	39.7	53.6	18.1	57.7	17.7	55.4	18.7	61.6	16.5	37.9	25.0	30.5	27.0

Table 3: Model performance on **baseline multiple-choice QA with no extra context**, listing both answer-aligned (ans) and hint-aligned (hint) selections. Each row shows results by category in English and Farsi, while the final row provides average accuracy scores across all categories. Higher "ans" counts indicate more correct answers, whereas "hint" counts reflect how often models choose an option related to a true-but-misleading hint, even though no hint was explicitly provided in this scenario.

Category	Qwen2.5-7B				GPT-4o				DeepSeek-V3				Claude-3.7				LLaMA-3.1-8B			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Literature	93	3	96	2	100	0	100	0	99	0	100	0	100	0	100	0	77	13	55	21
Education	91	3	92	3	96	2	98	2	100	0	99	1	100	0	99	1	79	12	66	13
Entertainment	100	0	97	2	100	0	99	0	100	0	99	0	99	1	99	0	78	14	57	23
Food	89	7	89	8	96	3	97	2	95	5	97	3	97	3	98	2	77	16	58	22
Geography	83	1	96	2	99	1	98	1	99	1	99	1	99	1	99	1	85	9	62	16
History	93	4	97	1	99	1	99	1	98	2	97	3	99	1	99	1	86	10	64	11
Holidays & Leisure	96	3	96	2	95	3	96	2	96	4	96	3	95	4	97	2	83	5	63	13
Religion	96	4	92	5	99	1	99	1	99	1	99	1	99	1	100	0	85	11	74	12
Science & Technology	91	3	96	2	98	2	99	1	99	0	99	1	98	2	99	0	81	8	57	14
Sports	93	1	100	0	100	0	99	0	100	0	98	0	100	0	100	0	84	8	62	15
Average Accuracy (%)	92.5	2.9	95.1	2.7	98.2	1.3	98.4	1.0	98.5	1.3	98.3	1.3	98.6	1.3	99.0	0.7	81.5	10.6	61.8	16.0

Table 4: Model performance on **multiple-choice QA with a correct explanation** provided in the prompt. Columns labeled "ans" represent how frequently each system chooses the correct answer, while "hint" indicates how often it instead selects a hint-related option. Results are broken down by category in English and Farsi, with the final row showing average accuracy scores across all categories.

QA + Explanation. Table 4 shows model performance when provided with a correct explanation aligned to the right answer. Accuracy often exceeds 90% across models in both Farsi and English, highlighting the strong effect of explicit factual support. Interestingly, models that previously labeled an explanation as Uncertain or even False in isolation still leverage it effectively once placed within the QA prompt. For example, DeepSeek-V3 in Farsi classifies only 57.1% of explanations as True (and 34.6% as Uncertain), yet surpasses 90% QA accuracy when those same explanations are provided in context. Qwen2.5-7B presents an interesting contrast, underperforming in English factual classification but still benefiting from the explanation during QA. LLaMA-3.1-8B again shows a language-based difference, gaining more from explanations in English than in Farsi. 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. We next compare the baseline results with those in Table 5, where prompts include a true-but-misleading hint. On average, the proportion of answers aligning with the hint rises from 21.5% to 63.5% in English and from 24.8% to 62.0% in Farsi, a gain of roughly 40 percentage points. Qwen2.5-7B exhibits the most drastic shift, a +60% jump in hint-based choices across both languages, despite frequently labeling hints as False in isolation. This disparity further corroborates that a model may not deem a statement true by itself, yet still be conditioned by it in a QA context. Claude-3.7 and DeepSeek-V3 also show substantial susceptibility to hints in Farsi, whereas GPT-4o

Category	Qwen2.5-7B				GPT-4o				DeepSeek-V3				Claude-3.7				LLaMA-3.1-8B			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Literature	8	83	7	85	45	39	45	43	19	75	28	63	27	62	32	60	27	50	29	37
Education	11	81	13	77	44	42	39	50	17	78	21	72	18	73	28	67	20	65	28	53
Entertainment	5	92	8	88	40	43	37	51	8	91	14	82	20	74	21	71	25	61	27	54
Food	17	72	19	73	38	45	32	49	19	73	23	72	25	67	28	60	27	52	25	47
Geography	13	75	11	86	45	38	43	43	17	72	20	68	23	63	31	59	19	63	26	45
History	13	78	11	83	52	33	51	42	25	66	38	58	32	60	43	52	25	57	21	45
Holidays & Leisure	10	86	5	87	31	59	27	57	11	83	21	74	19	74	27	68	28	50	27	34
Religion	23	75	14	77	53	37	50	40	25	72	35	59	33	60	41	54	31	61	27	53
Science & Technology	10	82	13	80	47	38	46	42	19	80	20	74	27	66	21	72	21	63	20	60
Sports	5	90	5	90	44	41	37	55	10	82	15	68	18	68	21	75	29	51	30	46
Average Accuracy (%)	11.5	81.2	10.6	82.6	43.9	40.1	40.7	47.2	18.0	76.0	23.7	67.0	24.2	66.1	31.3	66.6	25.5	54.3	26.2	49.3

Table 5: Model performance in **multiple-choice QA when true-but-misleading hints are explicitly provided**. Columns labeled "ans" indicate how frequently each system selects the correct answer, while "hint" shows how often it chooses the hint-based option. Results are grouped by category in English and Farsi, and the final row reports average accuracy scores across all categories.

and LLaMA-3.1-8B appear less prone to selecting the hint-based option outright, though their overall accuracy does decline significantly once a hint is introduced.

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. Category-level analyses reinforce these distinctions. Entertainment questions show the most pronounced accuracy drop under hint-based distractors, whereas History and Art & Literature prove less susceptible, likely because historical references appear more prominently in pretraining data. Explanations consistently boost accuracy across all categories (Table 4), highlighting how direct factual support stabilizes model performance even in the face of otherwise misleading content.

Discussion. Our findings emphasize two primary insights. First, LLMs still struggle with factual recognition in isolation, as reflected by True-label rates below 50% in factuality classification. Second, contextual framing strongly influence QA results: correct explanations often raise accuracy above 90%, whereas factual yet off-target hints can significantly reduce baseline performance or prompt models to select the hint-based option at high rates. Among the systems tested, GPT-4o, Claude-3.7, and DeepSeek-V3 manage these opposing forces more effectively in Farsi, whereas LLaMA-3.1-8B excels in English factual classification yet sometimes lags in Farsi QA. This language-dependent behavior underscores the importance of domain familiarity, Iranian content, in this case, which can bolster baseline accuracy but also influence how readily models adopt persuasive but ultimately irrelevant details.

6 Conclusion

We introduced a bilingual (Farsi–English) dataset aimed at examining LLMs vulnerability to factually correct yet misleading content in question-answering scenarios. As LLMs increasingly tackle ISQs, they must sift through diverse, sometimes distracting information that can undermine their responses. Our experiments reveal that LLMs are generally susceptible to additional details embedded in the prompt, regardless of the question's category or language. Intriguingly, these same systems may classify such true statements as False if presented out of context, suggesting that contextual framing plays a critical role in how models interpret information. Among the evaluated models, Qwen, DeepSeek, and Claude show greater susceptibility to persuasive-but-irrelevant content, while GPT shows relatively higher resilience, despite its accuracy also declining in the presence of misleading hints. Overall, our findings emphasize the importance of assessing how LLMs contend with accurate yet off-target details, especially in real-world question-answering contexts. Future efforts could explore enhanced context-filtering or verification strategies to mitigate the influence of true-but-misleading information and improve LLM 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.

Second, the overall size of our dataset is 1,000 Farsi samples, 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.

Third, our evaluation was conducted using a set of five models, which might also be considered as a few. This selection reflects a trade-off between breadth and depth; while we aimed to cover a diverse array of experimental conditions, doing so restricted our ability to include a larger number of models.

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*.

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*.

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*.

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, Francesco Pierri, and Debora Nozza. 2025. 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.

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 the table 8 while the target distributions are present in the table 7.

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 9. To collect the documents, we used the following category and sub-category definitions:

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 6: 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 7: 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, About the answer means that the hint is about the specific correct answer to the query.

Food

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

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 8: Distribution of counts across categories and subcategories

- **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., 2025). Accordingly, the table 10 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. The tables 12, 13, 14, 16, and 15 provide a full version of the tables that are presented in the paper for the five main tasks, with categorical values as well as the average values.

As mentioned in the paper, we also tested the most and least susceptible models to hints, based on the table 15, while providing both the explanation and the persuasive hints. 17 and 18 present the results of GPT4O on English and Farsi version of the resource. Moreover, 19 and 20 provide the same information for Qwen. As we can see, the results are close to the results of the experiments with questions and explanations. The reason, as we stated in the paper, is that explanations usually target the answer more directly compared to the hints that provide partial information. Accordingly, while presented together, the models might be able to answer correctly, when each of the hints or explanations is provided in isolation, the effect is different.

We provide the results of the models across the experiments and question classes. 21 and 22 are associated with the results of sonnet 3.7 on English and Farsi. Moreover, 23 and 24 are the DeepSeek' results, while 25 and 26 provide the GPT4o's results across question classes. Finally tables 29, 30, 27, and 28 provide Qwen and Llama's results on English and Farsi. The distribution of the classes of questions are provided in the table 6. We can see that, when hints are provided, there is a general drop in accuracies across all models and classes. However, there are also different patterns. For example, Deepseek shows a better performance for the Event class than the rest. Also, hints of different classes show different but close effectiveness

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 9: The prompt used for collecting Wikipedia documents.

rates. For example, for the English version and for Deepseek, Person, location, and other proper nouns reach higher effectiveness rates of above 80 percent.

Finally, we provide a comparative set of result tables for the two possible target values to explore the effectiveness of these two types of hints. Table 31 provides the results for the baseline. The table 32 provides the results for the case when the question and explanation are provided together. Finally, the 33 presents the QA + Hint setting results. I should note that only in the last table's experiment we actually provided the hints, and in the two other ones, we did not provide the hints.

The results of QA + Explanation + Hint are also provided in their respective tables, which are 17, 18, 19, and 20.

Based on the table 33, we can see that the hints that target a specific option, rather than providing distracting information about the answer, are more influential almost across all the tested models. Also, they are better in English than Farsi, while when the target is to answer, the effectiveness is larger in Farsi.

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 10: 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 11: The Farsi counterparts of the prompts provided in the table 1 for experiments.

	Qwen						GPT						DeepSeek						Claude						LLaMA						Average					
	exp			hint			exp			hint			exp			hint			exp			hint			exp			hint			exp			hint		
Category	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U
Arts & Lit.	44	56	0	48	42	10	58	20	22	67	14	19	61	3	36	71	4	25	51	15	34	44	16	40	25	73	2	28	70	2	47.8	33.4	18.8	51.6	29.2	19.2
Education	27	71	2	38	52	10	57	14	29	57	19	24	55	15	30	69	7	24	41	19	40	50	15	35	28	71	1	40	60	0	41.6	38.0	20.4	50.8	30.6	18.6
Ent.	33	63	4	48	48	4	62	12	26	66	18	16	47	7	46	65	7	28	19	21	60	36	15	49	33	67	0	30	69	1	38.8	34.0	27.2	49.0	31.4	19.6
Food	30	69	1	40	58	2	48	6	46	50	14	36	55	5	40	59	7	34	45	12	43	48	17	35	27	70	3	40	53	7	41.0	32.4	26.6	47.4	29.8	22.8
Geography	36	62	2	45	53	2	57	20	23	52	24	24	65	10	25	61	9	30	44	16	40	43	22	35	51	47	2	47	51	2	50.6	31.0	18.4	49.6	31.8	18.6
History	39	61	0	49	50	1	72	14	14	60	25	15	75	5	19	74	7	19	62	19	19	57	20	23	22	75	3	28	69	3	54.0	34.8	11.0	53.6	34.2	12.2
Holidays	21	75	4	41	56	3	52	12	36	47	14	39	48	5	47	54	5	41	30	12	58	43	16	41	25	73	2	34	65	1	35.2	35.4	29.4	43.8	31.2	25.0
Religion	34	64	2	39	60	1	49	16	35	60	14	26	55	7	38	74	2	24	43	15	42	50	15	35	37	61	2	36	61	3	43.6	32.6	23.8	51.8	30.4	17.8
Sci. & Tec	35	58	7	39	48	13	58	19	23	65	13	22	67	4	29	68	3	29	30	16	54	32	10	58	32	67	1	37	62	1	44.4	32.8	22.8	48.2	27.2	24.6
Sports	29	69	2	40	52	8	59	17	24	66	16	18	43	21	36	68	5	27	28	29	43	48	18	34	41	58	1	48	50	2	40.0	38.8	21.2	54.0	28.2	17.8
Avg	32.8	64.8	2.4	42.7	51.9	5.4	57.2	15.0	27.8	59.0	17.1	23.9	57.1	8.2	34.6	66.3	5.6	28.1	39.3	17.4	43.3	45.1	16.4	38.5	32.1	66.2	1.7	36.8	61.0	2.2	43.7	34.3	22.0	50.0	30.4	19.6

Table 12: Models' responses on Farsi explanations and hints while classifying them. T = True, F = False, U = Uncertain. **ans** is the count of answer matches. **hint** is the sum of hint option matches.

	Qwen						GPT						DeepSeek						Claude						LLaMA						Average					
	exp			hint			exp			hint			exp			hint			exp			hint			exp			hint			exp			hint		
Category	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U
Arts & Lit.	0	99	1	0	100	0	53	22	25	54	22	24	16	3	81	27	1	72	19	3	78	21	3	76	93	7	0	95	4	1	36.2	26.8	37.0	39.4	26.0	34.6
Education	0	100	0	0	99	1	45	34	21	53	24	23	16	4	80	30	9	61	16	7	77	28	7	65	86	13	1	91	7	2	32.6	31.6	35.8	40.4	29.2	30.4
Ent.	0	99	0	0	100	0	47	24	29	54	26	20	9	8	83	19	3	78	7	7	86	12	4	84	87	13	0	95	5	0	30.0	30.2	39.6	36.0	27.6	36.4
Food	0	100	0	0	100	0	40	19	41	44	22	34	30	4	66	34	5	61	24	6	70	30	10	60	95	2	3	93	4	3	37.8	26.2	36.0	40.2	28.2	31.6
Geography	0	100	0	0	100	0	52	24	24	46	27	27	30	6	64	38	2	60	23	4	73	27	6	67	92	6	2	93	4	3	39.4	28.0	32.6	40.8	27.8	31.4
History	1	99	0	0	100	0	69	24	7	63	25	12	45	5	50	52	5	43	41	7	52	44	8	48	96	4	0	90	7	3	50.4	27.8	21.8	49.8	29.0	21.2
Holidays	0	99	1	0	100	0	41	26	33	48	26	26	17	6	77	30	5	65	14	4	82	24	3	73	94	5	1	92	6	2	33.2	28.0	38.8	38.8	28.0	33.2
Religion	0	100	0	0	100	0	51	15	34	59	11	30	37	3	60	48	0	52	34	4	62	42	2	56	94	5	1	97	3	0	43.2	25.4	31.4	49.2	23.2	27.6
Sci. & Tec.	0	100	0	0	100	0	53	28	19	57	22	21	23	5	72	37	6	57	18	5	77	27	4	69	93	6	1	98	2	0	37.4	28.8	33.8	43.8	26.8	29.4
Sports	0	100	0	0	100	0	43	37	20	53	34	13	25	11	64	40	11	49	18	6	76	31	5	64	89	11	0	94	5	1	35.0	33.0	32.0	43.6	31.0	25.4
Avg	0.1	99.6	0.3	0	99.9	0.1	49.4	25.3	25.3	53.1	23.9	23	24.8	5.5	69.7	35.5	4.7	59.8	21.4	5.3	73.3	28.6	5.2	66.2	91.9	7.2	0.9	93.8	4.7	1.5	37.5	28.6	33.9	42.2	27.7	30.1

Table 13: Models' responses on English explanations and hints while classifying them. T = True, F = False, U = Uncertain. **ans** is the count of answer matches. **hint** is the sum of hint option matches.

	Qwen				GPT				DeepSeek				Claude				LLaMA				Avg			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
Category	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Lit.	36	26	31	20	59	16	45	37	56	17	66	14	60	20	71	13	41	23	22	28	50.4	20.4	47.0	22.4
Education	35	31	36	23	51	20	44	42	53	20	56	24	51	22	59	18	36	28	28	29	45.2	24.2	44.6	27.2
Ent.	32	25	34	24	55	19	40	43	51	21	53	19	55	17	57	15	31	28	27	29	44.8	22.0	42.2	26.0
Food	45	23	38	20	54	21	38	45	52	20	53	24	52	18	59	15	35	17	26	34	47.6	19.8	42.8	27.6
Geography	37	28	33	28	50	11	45	38	57	16	58	11	61	16	62	17	32	26	23	29	47.4	19.4	44.2	24.2
History	37	27	44	21	69	12	52	33	57	18	61	15	58	21	67	15	26	23	30	28	49.4	20.2	50.8	22.4
Holidays	31	30	34	25	46	21	31	59	42	23	49	20	55	15	57	25	32	27	29	21	41.2	23.2	40.0	30.0
Religion	41	34	40	26	66	12	53	37	59	15	68	16	66	15	68	10	42	28	38	28	55.2	20.8	53.4	23.4
Sci & Tec.	43	22	30	22	62	19	47	38	56	15	57	17	60	17	59	17	36	30	26	24	51.4	20.6	43.8	23.6
Sports	31	27	40	20	54	22	44	41	47	26	58	17	56	22	57	20	41	21	36	20	45.8	23.6	47.0	23.6
Avg	36.8	27.3	37.0	22.9	56.6	18.3	43.9	39.7	53.6	18.1	57.7	17.7	55.4	18.7	61.6	16.5	37.9	25.0	30.5	27.0	48.1	21.5	47.5	24.8

Table 14: The full comparison of models' answers matches, along with the average column, with the real answers and the persuasive hint options when the prompt includes no additional information other than the question and the options (**ans** is the count of answer matches; **hint** is the sum of hint option matches).

Category	Qwen				GPT				DeepSeek				Claude				LLaMA				Avg			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Lit.	8	83	7	85	45	39	45	43	19	75	28	63	27	62	32	60	27	50	29	37	25.2	61.8	28.2	57.6
Education	11	81	13	77	44	42	39	50	17	78	21	72	18	73	28	67	20	65	28	53	22.0	67.8	25.8	63.8
Ent.	5	92	8	88	40	43	37	51	8	91	14	82	20	74	21	71	25	61	27	54	19.6	72.2	21.4	69.2
Food	17	72	19	73	38	45	32	49	19	73	23	72	25	67	28	60	27	52	25	47	25.2	62.0	25.4	60.2
Geography	13	75	11	86	45	38	43	43	17	72	20	68	23	63	31	59	19	63	26	45	23.4	62.2	26.2	60.2
History	13	78	11	83	52	33	51	42	25	66	38	58	32	60	43	52	25	57	21	45	29.4	58.2	32.8	56.0
Holidays	10	86	5	87	31	59	27	57	11	83	21	74	19	74	27	68	28	50	27	34	19.8	70.4	21.4	64.0
Religion	23	75	14	77	53	37	50	40	25	72	35	59	33	60	41	54	31	61	27	53	33.4	65.0	33.4	56.6
Sci & Tec.	10	82	13	80	47	38	46	42	19	80	20	74	27	66	21	72	21	63	20	60	24.8	65.8	24.2	63.6
Sports	5	90	5	90	44	41	37	55	10	82	15	68	18	68	21	75	29	51	30	46	21.2	66.4	21.6	66.8
Avg	11.5	81.2	10.6	82.6	43.9	40.1	40.7	47.2	18.0	76.0	23.7	67.0	24.2	66.1	31.3	66.6	25.5	54.3	26.2	49.3	24.6	63.5	26.5	62.0

Table 15: Comparison of models' answers matches with the real answers and the persuasive hint options when the prompt includes persuasive hints, with the average values included in the table. **ans** is the count of answer matches. **hint** is the sum of hint option matches.

Category	Qwen				GPT				DeepSeek				Claude				LLaMA				Avg			
	English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi		English		Farsi	
	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint	ans	hint
Arts & Lit.	93	3	96	2	100	0	100	0	99	0	100	0	100	0	100	0	77	13	55	21	93.8	3.2	90.2	4.6
Education	91	3	92	3	96	2	98	2	100	0	99	1	100	0	99	1	79	12	66	13	93.2	3.4	90.8	4.0
Ent.	100	0	97	2	100	0	99	0	100	0	99	0	99	1	99	0	78	14	57	23	95.4	3.0	90.2	5.0
Food	89	7	89	8	96	3	97	2	95	5	97	3	97	3	98	2	77	16	58	22	90.8	6.8	87.8	7.4
Geography	83	1	96	2	99	1	98	1	99	1	99	1	99	1	99	1	85	9	62	16	93.0	2.6	90.8	4.2
History	93	4	97	1	99	1	99	1	98	2	97	3	99	1	99	1	86	10	64	11	95.0	3.6	91.2	3.4
Holidays	96	3	96	2	95	3	96	2	96	4	96	3	95	4	97	2	83	5	63	13	93.0	3.8	89.6	4.4
Religion	96	4	92	5	99	1	99	1	99	1	99	1	99	1	100	0	85	11	74	12	95.6	3.6	92.8	3.8
Sci & Tec.	91	3	96	2	98	2	99	1	99	0	99	1	98	2	99	0	81	8	57	14	93.4	3.0	90.0	3.6
Sports	93	1	100	0	100	0	99	0	100	0	98	0	100	0	100	0	84	8	62	15	95.4	1.8	91.8	3.0
Avg	92.5	2.9	95.1	2.7	98.2	1.3	98.4	1.0	98.5	1.3	98.3	1.3	98.6	1.3	99.0	0.7	81.5	10.6	61.8	16.0	93.9	3.5	90.5	4.3

Table 16: Comparison of models' answers matches with the real answers and the persuasive hint options when the prompt includes explanations, including the average values of models. **ans** is the count of answer matches. **hint** is the sum of hint option matches.

Catgory	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Matches cnt	Hint Matches %
Arts_&_Lit	99	100.00	3	0	0.00	96	0	0.00
Edu	98	98.00	3	0	0.00	97	1	1.03
Ent	100	100.00	1	0	0.00	99	0	0.00
FOOD	97	97.00	15	0	0.00	85	2	2.35
Geo	98	98.99	6	0	0.00	93	1	1.08
Hist	98	98.99	5	0	0.00	94	1	1.06
Holidays	94	95.92	3	3	100.00	95	0	0.00
Religion	98	98.00	4	0	0.00	96	1	1.04
Sci_&_Tec	98	98.99	4	0	0.00	95	1	1.05
SPORTS	99	99.00	0	0	0.00	100	0	0.00
Overall	979	98.49	44	3	6.82	950	7	0.74

Table 17: The results of GPT4O, while the explanation and hints are both included in the prompt on the English version. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint matches cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response.

Catgory	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Matches cnt	Hint Matches %
Arts_&_Lit	100	100.00	3	0	0.00	97	0	0.00
Edu	99	99.00	3	0	0.00	97	1	1.03
Ent	99	99.00	1	0	0.00	99	0	0.00
FOOD	97	97.00	15	0	0.00	85	2	2.35
Geo	98	98.00	6	0	0.00	94	1	1.06
Hist	99	99.00	5	0	0.00	95	1	1.05
Holidays	96	96.00	3	2	66.67	97	0	0.00
Religion	98	98.99	4	0	0.00	95	1	1.05
Sci_&_Tec	99	99.00	4	0	0.00	96	1	1.04
SPORTS	99	99.00	0	0	0.00	100	0	0.00
Overall	984	98.50	44	2	4.55	955	7	0.73

Table 18: The results of GPT4O, while the explanation and hints are both included in the prompt on the Farsi version. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint matches cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response.

Catgory	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Matches cnt	Hint Matches %
Arts_and_Lit	82	96.47	1	0	0.00	84	1	1.19
Edu	88	95.65	2	0	0.00	90	2	2.22
Ent	68	100.00	1	0	0.00	67	0	0.00
Food	83	89.25	13	4	30.77	80	4	5.00
Geo	95	95.96	5	0	0.00	94	3	3.19
Hist	89	92.71	4	0	0.00	92	5	5.43
Holidays	87	95.60	2	2	100.00	89	1	1.12
Religion	92	97.87	3	2	66.67	91	0	0.00
Sci_and_Tec	92	95.83	4	1	25.00	92	1	1.09
Sports	86	100.00	0	0	0.00	86	0	0.00
Overall	862	95.78	35	9	25.71	865	17	1.97

Table 19: The results of Qwen, while the explanation and hints are both included in the prompt on the English version. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint matches cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response.

Catgory	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Matches cnt	Hint Matches %
Arts_and_Lit	97	97.00	3	1	33.33	97	0	0.00
Edu	93	93.94	3	1	33.33	96	1	1.04
Ent	97	97.98	1	0	0.00	98	1	1.02
FOOD	89	89.90	15	6	40.00	84	2	2.38
Geo	94	94.00	6	2	33.33	94	1	1.06
Hist	94	94.00	5	0	0.00	95	4	4.21
Holidays	94	94.00	3	2	66.67	97	0	0.00
Religion	93	93.00	4	0	0.00	96	4	4.17
Sci_and_Tec	94	94.00	4	1	25.00	96	1	1.04
SPORTS	98	98.99	0	0	0.00	99	0	0.00
Overall	943	94.68	44	13	29.55	952	14	1.47

Table 20: The results of Qwen, while the explanation and hints are both included in the prompt on the Farsi version. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint matches cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	149	40	26.85	3	0	0.00	146	99	67.81
Location	177	40	22.60	10	0	0.00	167	125	74.85
OPN	108	33	30.56	17	6	35.29	91	63	69.23
Other	42	10	23.81	3	1	33.33	39	30	76.92
Date/Time	196	36	18.37	2	0	0.00	194	142	73.20
Artwork	25	8	32.00	2	2	100.00	23	14	60.87
Group/Org	103	34	33.01	5	0	0.00	98	65	66.33
Number	159	27	16.98	0	0	0.00	159	108	67.92
Event	27	14	51.85	2	1	50.00	25	11	44.00

Table 21: The results of Claude for different question classes, while the hints were included in the prompt on the English version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	150	55	36.67	3	2	66.67	147	86	58.50
Location	178	51	28.65	10	2	20.00	168	115	68.45
OPN	108	39	36.11	17	6	35.29	91	57	62.64
Other	42	16	38.10	3	0	0.00	39	25	64.10
Date/Time	198	39	19.70	2	0	0.00	196	148	75.51
Artwork	25	9	36.00	2	2	100.00	23	12	52.17
Group/Org	102	36	35.29	5	1	20.00	97	61	62.89
Number	160	33	20.63	0	0	0.00	160	110	68.75
Event	27	15	55.56	2	1	50.00	25	10	40.00

Table 22: The results of Claude for different question classes, while the hints were included in the prompt on the Farsi version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	25	16.56	3	1	33.33	148	119	80.41
Location	178	31	17.42	10	1	10.00	168	140	83.33
OPN	109	27	24.77	17	5	29.41	92	74	80.43
Other	42	11	26.19	3	0	0.00	39	28	71.79
Date/Time	198	24	12.12	2	0	0.00	196	163	83.16
Artwork	25	3	12.00	2	1	50.00	23	21	91.30
Group/Org	105	17	16.19	5	0	0.00	100	85	85.00
Number	163	25	15.34	0	0	0.00	163	115	70.55
Event	27	7	25.93	2	1	50.00	25	18	72.00

Table 23: The results of Deepseek for different question classes, while the hints were included in the prompt on the English version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	38	25.17	3	1	33.33	148	102	68.92
Location	178	42	23.60	10	0	0.00	168	127	75.60
OPN	109	37	33.94	17	5	29.41	92	61	66.30
Other	42	11	26.19	3	0	0.00	39	29	74.36
Date/Time	198	35	17.68	2	1	50.00	196	147	75.00
Artwork	25	8	32.00	2	0	0.00	23	17	73.91
Group/Org	105	24	22.86	5	2	40.00	100	74	74.00
Number	165	28	16.97	0	0	0.00	165	120	72.73
Event	27	12	44.44	2	1	50.00	25	13	52.00

Table 24: The results of Deepseek for different question classes, while the hints were included in the prompt on the Farsi version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	72	47.68	3	3	100.00	148	59	39.86
Location	178	78	43.82	10	0	0.00	168	84	50.00
OPN	109	46	42.20	17	7	41.18	92	44	47.83
Other	42	19	45.24	3	1	33.33	39	17	43.59
Date/Time	197	82	41.62	2	1	50.00	195	76	38.97
Artwork	25	11	44.00	2	2	100.00	23	9	39.13
Group/Org	104	59	56.73	5	1	20.00	99	39	39.39
Number	165	55	33.33	0	0	0.00	165	65	39.39
Event	27	17	62.96	2	1	50.00	25	6	24.00

Table 25: The results of GPT for different question classes, while the hints were included in the prompt on the English version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	65	43.05	3	3	100.00	148	69	46.62
Location	178	73	41.01	10	2	20.00	168	86	51.19
OPN	109	44	40.37	17	7	41.18	92	48	52.17
Other	42	19	45.24	3	0	0.00	39	22	56.41
Date/Time	196	75	38.27	2	1	50.00	194	92	47.42
Artwork	25	12	48.00	2	2	100.00	23	9	39.13
Group/Org	105	51	48.57	5	2	40.00	100	45	45.00
Number	164	54	32.93	0	0	0.00	164	73	44.51
Event	27	14	51.85	2	1	50.00	25	10	40.00

Table 26: The results of GPT for different question classes, while the hints were included in the prompt on the Farsi version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	42	27.81	3	2	66.67	148	81	54.73
Location	177	46	25.99	10	9	90.00	167	97	58.08
OPN	109	34	31.19	17	15	88.24	92	39	42.39
Other	42	15	35.71	3	1	33.33	39	17	43.59
Date/Time	191	45	23.56	2	2	100.00	189	110	58.20
Artwork	25	4	16.00	2	0	0.00	23	16	69.57
Group/Org	105	36	34.29	5	4	80.00	100	55	55.00
Number	158	23	14.56	0	0	0.00	158	105	66.46
Event	27	7	25.93	2	2	100.00	25	14	56.00

Table 27: The results of Llama for different question classes, while the hints were included in the prompt on the English version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	150	34	22.67	3	3	100.00	147	80	54.42
Location	178	44	24.72	10	7	70.00	168	84	50.00
OPN	109	38	34.86	17	12	70.59	92	29	31.52
Other	40	13	32.50	3	3	100.00	37	15	40.54
Date/Time	196	56	28.57	2	2	100.00	194	68	35.05
Artwork	25	8	32.00	2	0	0.00	23	12	52.17
Group/Org	104	24	23.08	5	3	60.00	99	46	46.46
Number	164	34	20.73	0	0	0.00	164	77	46.95
Event	27	9	33.33	2	1	50.00	25	14	56.00

Table 28: The results of Llama for different question classes, while the hints were included in the prompt on the Farsi version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	20	13.25	3	0	0.00	148	125	84.46
Location	178	23	12.92	10	1	10.00	168	147	87.50
OPN	109	25	22.94	17	7	41.18	92	66	71.74
Other	42	11	26.19	3	2	66.67	39	25	64.10
Date/Time	197	5	2.54	2	1	50.00	195	182	93.33
Artwork	25	2	8.00	2	1	50.00	23	22	95.65
Group/Org	105	12	11.43	5	2	40.00	100	85	85.00
Number	153	10	6.54	0	0	0.00	153	129	84.31
Event	27	7	25.93	2	1	50.00	25	18	72.00

Table 29: The results of Qwen for different question classes, while the hints were included in the prompt on the English version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Catgory	Tot	Acc. cnt	Acc%	Targ=1	Diff cnt	Diff %	Targ=0	Hint Mat cnt	Hint Mat %
Person	151	20	13.25	3	0	0.00	148	125	84.46
Location	178	23	12.92	10	1	10.00	168	147	87.50
OPN	109	25	22.94	17	7	41.18	92	66	71.74
Other	42	11	26.19	3	2	66.67	39	25	64.10
Date/Time	197	5	2.54	2	1	50.00	195	182	93.33
Artwork	25	2	8.00	2	1	50.00	23	22	95.65
Group/Org	105	12	11.43	5	2	40.00	100	85	85.00
Number	153	10	6.54	0	0	0.00	153	129	84.31
Event	27	7	25.93	2	1	50.00	25	18	72.00

Table 30: The results of Qwen for different question classes, while the hints were included in the prompt on the Farsi version. Tot is the total number of cases when the model provided a meaningful answer. Acc. Cnt means the number of true answers, which match the option targeted by the explanation. Acc. % is the percentage of the previous column. Targ=1 is the number of samples where their hints target no specific option other than the answer itself, while Target=0 is the number of samples that their hint actually target a wrong option. Diff cnt is then the number of times where the model answered incorrectly when Target=1, and Diff % is its percentage. Hint mat cnt and % are also the number of times and percentages where the model's response matches the hint option. Note: We only calculate the numbers based on the cases where the model actually provided a response. OPN= Other proper and common nouns.

Model	English (target=1)	English (target=0)	Farsi (target=1)	Farsi (target=0)
Claude	34.09	19.13	31.82	16.58
Deepseek	52.27	19.67	40.91	17.15
GPT	31.82	17.30	36.36	41.82
Llama	68.18	27.16	86.11	29.59
Qwen	56.82	26.33	52.27	21.80

Table 31: The results of our five tested models for the two values of the target field, when the prompt did not include any additional information. Target=1 means that the hint was not about a wrong option, but it was about the correct answer. Target=0 means that the hint was about one of the wrong options.

Model	English (target=1)	English (target=0)	Farsi (target=1)	Farsi (target=0)
Claude	11.36	0.84	4.55	0.52
Deepseek	13.64	0.73	6.82	1.05
GPT	6.82	1.05	4.55	0.84
Llama	60.47	8.45	77.27	13.22
Qwen	38.10	1.40	40.91	0.94

Table 32: The results of our five tested models for the two values of the target field, when the prompt included explanations. Target=1 means that the hint was not about a wrong option, but it was about the correct answer. Target=0 means that the hint was about one of the wrong options.

Model	English (target=1)	English (target=0)	Farsi (target=1)	Farsi (target=0)
Claude	22.73	69.75	31.82	65.96
Deepseek	20.45	79.98	22.73	72.18
GPT	36.36	41.82	40.91	47.64
Llama	79.55	56.75	70.45	44.78
Qwen	34.09	84.73	11.36	85.97

Table 33: The results of our five tested models for the two values of the target field, when the prompt included persuasive hints. Target=1 means that the hint was not about a wrong option, but it was about the correct answer. Target=0 means that the hint was about one of the wrong options.