

FarSense: A Comprehensive Commonsense Benchmark and Evaluation Framework for the Farsi Language

Kamyar Zeinalipour¹, Neda Jamshidi¹, Seyedehbahareh Hejazi¹,
Marco Maggini¹, Monica Bianchini¹, Simone Paoletti¹, Marco Gori¹

¹University of Siena, Via Roma 56, 53100 Siena SI, Italia

Correspondence: kamyar.zeinalipour2@unisi.it

Abstract

Although Farsi is widely spoken, no comprehensive benchmark exists for assessing commonsense reasoning in language models. We therefore present **FarSense**, a 6-task benchmark for Farsi covering True/False judgment, multiple-choice questions, Explanation, Cause-Effect inference, Counterfactual reasoning, and Knowledge Completion. Starting from Farsi-Wikipedia, we filtered noise and retained 4,210 passages, rewrote them into realistic daily scenarios, and derived the above tasks from each scenario. Scenario and task generation quality was first judged via native-speaker annotations on outputs from five major LLMs—GPT-4o, Gemini-2.5-Flash, Mistral-Large, Qwen-Plus, and DeepSeek-Chat. Gemini-2.5-Flash demonstrated the highest performance, leading to its use in generating a large-scale dataset, subsequently finalized through meticulous two-step human validation. Using **FarSense**, we measured the commonsense ability of the same five flagship LLMs and also fine-tuned six compact models (1B–24B parameters) before re-evaluating them. To ensure broad applicability, task wording was designed to minimize dialectal, cultural, or religious bias. Experiments show that targeted fine-tuning yields substantial gains, confirming **FarSense** as a reliable, openly licensed resource for advancing reproducible commonsense understanding research in Farsi NLP. We publicly release all code and data at <https://github.com/KamyarZeinalipour/FarSense>.

1 Introduction

Although Farsi is widely spoken, there is currently no comprehensive benchmark for evaluating commonsense reasoning in language models for this language. This notable gap prevents an accurate assessment of how effectively state-of-the-art language models can handle commonsense reasoning tasks in Farsi (Khashabi et al., 2021). Existing

Task 1: Multiple Choice Question

سرا از چه نظر گریه ناشی از تماشای فیلم غمگین را با گریه ناشی از خرد کردن پیار متفاوت می‌داند؟
Q: In what way does Sara consider crying from watching a sad movie different from crying while chopping onions?

وجود یا عدم وجود احساسات عمیق پشت آن
Correct Answer: B: The presence or absence of deep emotions behind it.

Task 2: True/False judgment

گریه کردن نقصانه در مواجهه با اتفاقات ناخوشاید رخ مردود
Statement: Crying only happens in response to unpleasant events.

Answer: False

Task 3: Cause-Effect inference

مغز انسان تواند برای مسئلک پیش امده توجهی پیدا کند
Cause: The human brain cannot find a justification for a problem that has occurred

اصفات فرد تعلیم شود
Effect: The person's emotions are released

Task 4: Completion

شکهایی که هنگام خرد کردن پیاز تولید می‌شود، صرفاً یک اکشن شیمیایی است و «—» ندارند.
Sentence: The tears produced while chopping onions are merely a chemical reaction and lack "—".

Answer: احساس
Answer: emotion

Task 5: Explanation

چرا سارا فکر می‌کند که همای گریه می‌کند، ممکن است با مشکلات عمیق‌تری روبرو باشد؟
Q: Why does Sara think that her friend who cries constantly might be facing deeper problems?

سرا معتقد است گریه زیادی اتفاق می‌افتد که مغز تواند برای مشکلات راه حلی پیدا کند گریه می‌داند از شکایاتی مغز او در این قیمت توجهی برای امکان برای مشکلات است.
A: Sara believes that crying happens when the brain cannot find a solution to problems. Her friend's constant crying is a sign that his brain is unable to justify or cope with his issues.

Task 6: Counterfactual reasoning

اگر سارا به جای فیلم غمگین، یک کمدی پیار خندهدار تماشا کرده بود، ای ممکن بود زیرا هم اشک بربری؟
Q: If Sara had watched a very funny comedy instead of a sad movie, could she still have cried?

بله، پیار ممکن بود. سارا ادر پایان متن اشاره می‌کند که گریه می‌تواند ناشی از شادی پیدا کند و حسر نیز پاشد و ریشه‌هایی بازد.
A: Yes, it is quite possible. Sara mentions at the end that crying can also result from overwhelming joy and deep roots.

Figure 1: Illustrative example of subtasks. Full subtasks are provided in Appendix A

benchmarks primarily focus on English, including datasets such as CommonsenseQA (Talmor et al., 2019), COPA (Roemmele et al., 2011), Social IQa (Sap et al., 2019), and SemEval-2020 Task 4 (Wang et al., 2020). However, such resources are scarce or entirely absent for most other languages, including Farsi, highlighting a broader multilingual limitation in commonsense research (Sakai et al., 2024b; Shamsfard, 2019). Prior work on Farsi has typically addressed general language understanding tasks (e.g., ParsiNLU; Khashabi et al., 2021) or specialized domains (e.g., Farsi social norms; Safafari et al., 2024), but has not tackled the broader and crucial area of commonsense reasoning comprehensively.

Research Questions. This study addresses the overarching question:

Problem Statement

How can we leverage Large Language Models to effectively generate and utilize a comprehensive benchmark dataset for evaluating and enhancing commonsense reasoning capabilities in Farsi?

Specifically, we explore four research questions:

- **RQ1:** How effective are current state-of-the-art LLMs for generating high-quality, diverse commonsense reasoning tasks and scenarios in Farsi?
- **RQ2:** RQ2: How effective is a hybrid data-generation approach—combining LLM output and rigorous human validation—for developing a comprehensive Farsi commonsense benchmark?
- **RQ3:** What is the performance of leading large language models (LLMs) on various commonsense reasoning tasks in Farsi, and which types of commonsense queries pose the greatest challenge?
- **RQ4:** How much does fine-tuning smaller language models on the **FarSense** dataset improve Farsi commonsense reasoning compared to (i) their own off-the-shelf (pre-tuning) baselines and (ii) larger general-purpose models?

Task framing: FarSense evaluates *context-grounded commonsense reasoning* (CG-CSR): models must apply causal, temporal, social, and counterfactual schemas to a **provided narrative**. This contrasts with (i) **knowledge-centric CSR** (e.g., CommonsenseQA(Talmor et al., 2019), HellaSwag(Zellers et al., 2019)), which probes unstated world knowledge, and (ii) **extractive reading comprehension** (e.g., SQuAD(Rajpurkar et al., 2016)/TriviaQA(Joshi et al., 2017)), which can often be solved by span lookup. In FarSense, answers are *non-extractive* (no-span rule) and require reasoning over the given scenario, not recall of external facts.

Wikipedia text

گریه و اشکی حسماً است که با جاری شدن اشک و تغییر حالت چهره همراه می‌شود و با اشک‌های ناشی از مرگ و میری یا شیوه‌های متفاوت است. اغلب، گریه در واقعیت به ناشی از درک یا توجه یک مشکل حسماً پذید می‌آید و می‌تواند از آن‌دوهه افسرگی یا حتی شادی شنیده ناشی شود...

Crying is an emotional response accompanied by tears and changes in facial expression, and it differs from tears caused by chemical stimuli. Often, crying arises when the mind is unable to understand or rationalize an emotional problem, and it may stem from sorrow, depression, or even intense joy...

Scenario

سara دشیب با دین پک فیلم غول‌گیر پی‌لختگیر گردید که اشک‌هایش همراه با لرزش ایل و تله‌های آرام سرازد شد. او به تفاوت میان اشک حسماً و اشک ناشی از پیاره کرد و پنهان گردید، راهیست برای خلیعی احساسات و قیچی دن پاسخی برای درد ندارد...

Last night, Sara cried uncontrollably while watching a sad movie; her tears streamed down along with trembling lips and quiet sobs. She thought about the difference between emotional tears and those caused by onions, and realized that crying is a way to release emotions when the mind has no answer for the pain...

Figure 2: Illustrative example of generated scenario. Full texts are provided in Appendix A.

Approach and Results. We introduce **FarSense**, a robust benchmark comprising six distinct commonsense subtasks for Farsi. We initially gathered approximately 4,210 paragraphs from Farsi Wikipedia, converting them into realistic everyday scenarios enriched with commonsense elements (Figure 2). Scenarios and associated tasks were initially generated using five major LLMs (GPT-4o, Gemini-2.5-Flash, Mistral Large, Qwen-Plus, and DeepSeek-Chat), evaluated through rigorous human annotation. Gemini-2.5-Flash emerged as the best-performing model and was subsequently used to generate the complete benchmark, undergoing meticulous two-step human validation. The subtasks include verifying factual statements (True/False), Explanation, multiple-choice questions, cause-effect reasoning, counterfactual reasoning, and knowledge completion tasks (Figure 1). Our large-scale evaluation reveals a wide performance band (overall macro scores: 0.663–0.700) across five flagship LLMs: while binary facts and multiple-choice items are near-solved (in the 0.95–0.97 accuracy range), causal and counterfactual generation remain starkly difficult ($\text{BLEU} \leq 0.11$). Crucially, fine-tuning six compact models (1 B–24 B) on **FarSense** lifts their macro score by an average of +0.10 ($\approx 10 \text{ pp}$)—and by +23 pp for *Llama-3-1B*—confirming that the benchmark is not only diagnostic but also an effective training curriculum for resource-constrained Persian models.

Contributions. Our main contributions include: (1) Introducing **FarSense**, the first comprehensive commonsense reasoning benchmark specifically designed for Farsi, consisting of six diverse, human-validated subtasks; (2) Proposing and validating an effective hybrid methodology combining LLM-

driven scenario generation with rigorous human annotation; (3) Conducting extensive evaluations of multiple state-of-the-art language models in Farsi commonsense reasoning, providing insights into their capabilities and limitations; (4) Demonstrating the effectiveness of fine-tuning using **FarSense** to significantly improve commonsense reasoning performance in smaller-scale Farsi models. (5) Publicly releasing the **FarSense** benchmark dataset, associated source code, and fine-tuned models to facilitate reproducibility and encourage further research on Farsi commonsense reasoning.

Paper Structure. Section 2 reviews related work, Section 3 details dataset creation, Section 4 reports experiments, Section 5 concludes, and Section 6 discusses the limitations of the research.

2 Related Work

Commonsense Reasoning Benchmarks in English: Most commonsense reasoning benchmarks in NLP have focused on English, leading to a wide range of datasets targeting different types of reasoning. **CommonsenseQA** (Talmor et al., 2019) introduced multiple-choice questions based on ConceptNet to evaluate general commonsense knowledge. **COPA** (Roemmele et al., 2011) focuses on causal reasoning, while **Social IQa** (Sap et al., 2019) explores social motivations and interactions. SemEval-2020 Task 4 (Wang et al., 2020) evaluates both validation and explanation of commonsense knowledge. While these benchmarks have driven substantial progress, they remain limited to English and do not capture the linguistic and cultural nuances needed for multilingual commonsense evaluation.

Multilingual and Low-Resource Commonsense Datasets: To broaden the scope beyond English, several multilingual datasets have emerged. Translation-based resources like **XCOPA** (Ponti et al., 2020), an extension of COPA, offer initial steps toward multilingual commonsense reasoning. However, such datasets often inherit cultural assumptions and biases from English, limiting their validity across languages (Shamsfard, 2019). More recent benchmarks such as **mCSQA** (Sakai et al., 2024a) use hybrid approaches that combine LLM-generated content with human validation to create culturally relevant and cost-effective evaluations. Likewise, **NormBank** (Ziems et al., 2023) leveraged LLMs and human refinement to construct

a large-scale commonsense knowledge base. Inspired by these methods, **FarSense** adopts a similar hybrid strategy to ensure culturally grounded and high-quality benchmarks tailored to Farsi.

LLMs as Generators and Evaluators: Recent work leverages large language models in low-resource pipelines along two complementary axes: (i) *generation* of supervision and evaluation material, and (ii) *solver/evaluator* use. On the generation side, LLMs automate quiz authoring in Turkish (Zeinalipour et al., 2024b) and underpin PersianMCQ-Instruct for multiple-choice question creation in Farsi (Zeinalipour et al., 2025a). Puzzle-centric resources similarly scale via LLMs, producing crosswords and clue corpora in Italian, Turkish, and Arabic (Zeinalipour et al., 2024a,c, 2025b,d), with data-driven methods (e.g., Clue-Instruct) improving clue quality (Zugarini et al., 2024b). As evaluators/solvers, LLMs answer crossword clues (ECWCA) (Zugarini et al., 2024a), deliver rubric-aligned feedback on student writing (Zeinalipour et al., 2024d), and help construct targeted benchmarks in Arabic, from commonsense reasoning (Lamsiyah et al., 2025) to multi-dialect dialogue via *Shawarma Chats*, a triple-dialect benchmark (Egyptian, Maghrebi, MSA) with evaluated with automatic metrics (Zeinalipour et al., 2025c). These trends motivate our CG-CSR design for **FarSense**: LLM-based *Scenario Rewriting* and *Task Derivation* under a strict no-span rule, followed by two-stage native-speaker validation for cultural grounding (Section 3) and downstream solver assessment (Section 4).

Farsi Language Resources and the Commonsense Gap: Despite Farsi’s wide usage, few resources specifically address commonsense reasoning. Datasets like **ParsiNLU** (Khashabi et al., 2021) focus on general NLP tasks, while **PSN** (Safari et al., 2024) targets Farsi social norms. These lack coverage of general, everyday commonsense scenarios. The **Khayyam Challenge** (Ghahroodi et al., 2024) provides an extensive evaluation of reasoning skills across disciplines but does not target commonsense specifically. Similarly, **FarsEval-PKBETS** (Shamsfard et al., 2024) introduces a diverse set of knowledge-intensive tasks, yet leaves room for improvement in model performance on practical reasoning. None of these datasets provide a comprehensive, scenario-based commonsense benchmark for Farsi.

FarSense addresses this gap by constructing

	CG-CSR (ours)	CSR (knowledge)	RC (extractive)
Input basis	Provided narrative	World knowledge	Provided passage
Answerability	Closed-world	Open-world	Closed-world
Answer form	Non-extractive	Option/short fact	Often extractive span
Primary skills	Causal/temporal/social/counterfactual	Conceptual recall	Span/local inference
Examples	FarSense	CSQA/HellaSwag	SQuAD/TriviaQA

Table 1: Positioning FarSense within reasoning benchmarks.

culturally contextualized commonsense tasks directly from Farsi Wikipedia and realistic daily settings, thereby avoiding translation artifacts. Its two-stage design—LLM generation followed by expert human validation—follows successful practices from multilingual resources (Sakai et al., 2024a; Ziems et al., 2023). As the first dedicated Farsi commonsense benchmark, **FarSense** not only fills a critical void in the resource landscape but also offers a replicable methodology for future benchmarks in other low-resource languages.

Positioning: As summarized in Table 1, we distinguish three families of benchmarks. (1) **CSR (knowledge-centric)** targets implicit world knowledge and associative plausibility, often without a supporting passage (e.g., CommonsenseQA, HellaSwag; (Talmor et al., 2019; Zellers et al., 2019)). (2) **RC (extractive)** evaluates span or short-answer retrieval from a given passage (e.g., SQuAD, TriviaQA; (Rajpurkar et al., 2016; Joshi et al., 2017)). (3) **CG-CSR (ours)** requires *non-extractive* reasoning over a provided narrative with an explicit no-span constraint and skills like causal, temporal, social, and counterfactual inference. CG-CSR is closest to non-extractive RC that demands reasoning beyond span lookup (e.g., DROP, NarrativeQA; (Dua et al., 2019; Kočiský et al., 2018)), but differs in its emphasis on commonsense schemas rather than reading-comprehension objectives.

3 Dataset Construction

Our goal was to build a *native* Farsi commonsense benchmark that (i) covers a broad range of reasoning types, (ii) is free from translation artefacts, (iii) is large enough to fine-tune small to medium-sized models. Detailed steps are enumerated below.

3.1 Source Collection

We deliberately restrict our crawl to important portions of Farsi Wikipedia.

Starting from ten community-maintained indices—such as *Featured Articles*, *Vital Articles*,

and *Most-Viewed by Topic*—spanning mathematics, history, biology, and literature, ... we retrieve the pages referenced by each list (Wikipedia contributors, 2025b,a,d,c).

This yields **8,894** candidate articles. From each, we retain only the lead paragraph—typically a concise, self-contained summary—then apply three automatic screens: (i) Paragraphs shorter than 100 words are discarded to avoid stubs and disambiguation pages; (ii) Pages containing sensitive or disallowed content are excluded in line with our ethics protocol. (iii) Navigation or list pages are filtered out via namespace tags and heuristic detection of excessive bullet usage. After filtering, **4,210** high-quality seed paragraphs remain, which form the input to the scenario re-writing stage in the section (3.2).

3.2 Scenario Rewriting

Our pipeline separates content sourcing from commonsense packaging. Starting from 4,210 high-quality lead paragraphs extracted from pedagogically valuable portions of Farsi Wikipedia (Section 3.1), we automatically produce realistic, everyday narratives as follows:

Prompt engineering: A seed COT prompt was iteratively refined on 30 pilot passages using manual error analysis; the final form (B) yields consistent output across five flagship LLMs—GPT-4o (OpenAI et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), Qwen-Plus (Bai et al., 2023), Mistral-Large (Mistral AI, 2025), and DeepSeek-Chat (DeepSeek-AI et al., 2025).

Automatic generation: Each passage was submitted to five flagship LLMs to generate a 100–300 word commonsense scenario in Farsi that implicitly conveys all source facts while avoiding specialist terminology.

Detox and sanity checks: We applied regex filters to remove profanity, and verified that the output length met roughly the 100–300 word target. Scenarios failing any check were regenerated.

3.3 Task Derivation

Using the same prompt-engineering strategy outlined in Section 3.2, we crafted a *second* generation template (Appendix B). For every validated scenario, we invoked the same LLMs—GPT-4o, Gemini-2.5-Flash, Qwen-Plus, Mistral-Large, and DeepSeek-Chat—with this second template (Appendix B) to create six distinct commonsense items, including one multiple-choice question, two True/False statements, a Cause→Effect pair, a cloze completion item, an explanation prompt–answer pair, and a counterfactual question–answer pair. The template enforces three quality gates: (i) the no-span rule (answers must not appear verbatim in the scenario), (ii) explicit tagging of the underlying commonsense skill (`cs_skill` in *physical*, *social*, *temporal*, *intentional*, *causal*, *deontic*), and (iii) a single-sentence justification why the generated task is commonsense that could help the human annotators (`why_cs`).

3.4 Model Selection

To identify the most reliable LLM for large-scale scenario and task generation, we ran a controlled pilot on **100** randomly chosen Wikipedia lead paragraphs (Section 3.1). Each paragraph was processed twice—once with the *Scenario Re-writing* prompt and once with the *Task Derivation* prompt—using the five flagship models already introduced. This yielded **500** candidate scenarios and **3,500** task items (7 per scenario¹).

Human evaluation protocol. Two native-Farsi annotators, who grew up in Iran and currently live in Italy,² with substantial experience in commonsense reasoning—both holding a Master’s degree—individually rated each scenario and its derived tasks using a three-level scale: **A** (acceptable), **B** (partially acceptable), **C** (unacceptable). Guidelines and examples (see Appendix C and D) were provided; annotators were blind to the source model.

Inter-annotator reliability. During the pilot annotation phase, the two annotators achieved **94% raw agreement**. However, Cohen’s $\kappa = 0.35$ appears comparatively low, reflecting the well-known *prevalence paradox* (Byrt et al., 1993), where an uneven distribution of categories inflates the expected

¹Two distinct True/False statements were generated to cover both truth values.

²Both annotators are authors of this paper and participated voluntarily without receiving any payment.

Model	Scenario Score ↑	Tasks Score ↑
Gemini-2.5-Flash	2.6	2.3
GPT-4o	2.4	2.2
Mistral-Large	2.3	2.1
Qwen-Plus	2.2	2.0
DeepSeek-Chat	2.1	1.9

Table 2: Average human-judged quality (3=best) on the 100-paragraph pilot.

by-chance term and thereby deflates κ even under high observed agreement. All discrepancies were subsequently reviewed and resolved through consensus discussion, and the **released dataset** consequently exhibits **100% agreement**. Detailed agreement scores are provided in Table 10 in the Appendix 10. After independent scoring, all conflicting labels were jointly discussed until a consensus.

Scoring. Labels were mapped to numerical scores (A=3, B=2, C=1) and averaged per model for (i) scenario quality and (ii) the aggregated quality of all associated tasks. Results are summarised in Table 2.

Decision. Gemini-2.5-Flash achieved the highest mean scores on both dimensions and was therefore chosen as the backbone model for generating the full **FarSense** benchmark. Subsequent human validation ensured the same rigor and quality applied at scale.

3.4.1 Large-scale Generation

After selecting **Gemini-2.5-Flash** as the generator model (Section 3.4), we applied it to the full set of 4,210 Wikipedia lead paragraphs, generating both (i) rewritten scenarios and (ii) six associated commonsense tasks for each paragraph (Section 3.3). This procedure produced **4,210** scenario–task bundles (*i.e.*, 29,470 individual items).

Two stage human validation. Each bundle subsequently underwent a rigorous two-stage review:

1. **First pass (edit or regenerate).** Annotator A inspected every machine-generated bundle, performing minimal edits (grammar, spelling, answer-key alignment) or fully regenerating items that contained unrecoverable issues. Examples are provided in Appendix F.
2. **Second pass (consensus).** Annotator B independently reviewed the revised bundle. When-

ever disagreement arose, the annotators discussed the item until consensus was reached.

Of the 4,210 bundles, **2,531** (60.1%) required at least one edit, and **753** (17.9%) were fully regenerated. The resulting corpus constitutes the final **FarSense** dataset used in Section 4.

3.5 Dataset Statistics

Table 3 summarizes the quantitative profile of the **FarSense** corpus. In total, the benchmark comprises **29,470** annotated items—six subtasks (with two True/False items), i.e., seven items per scenario (3.3). The entire corpus contains **12.9 M** tokens, with an average of **438** tokens and a median of **429** tokens per item. Lexical diversity, measured by the *type–token ratio* (TTR), averages **0.54** and is broadly consistent across tasks (0.52–0.56), indicating that none of the subtasks is dominated by formulaic wording.

Task-level observations.

- **Length variation.** Multiple-choice questions are the longest (≈ 481 tokens on average) because they include the prompt, four answer options, and detailed rationales. True/False items are the shortest (≈ 413 tokens) but exhibit the highest average TTR (0.56), reflecting the strong lexical variety introduced by affirmative vs. negative statements.
- **Lexical diversity.** Across the corpus, we observe 71,347 unique word types—about one distinct type per 181 tokens—which is in line with benchmarks of similar scale in English. The low overall corpus TTR (0.01) results from the large sample size and is typical of datasets exceeding 10 M tokens.
- **Balanced coverage.** The seven item types each contribute roughly 13.5–15.7% of the total token count, so no single type dominates the training signal when the benchmark is used for fine-tuning (4.3).

These statistics confirm that **FarSense** offers both breadth—in terms of reasoning skills—and depth—in terms of lexical variety—making it a suitable benchmark for evaluating and improving commonsense reasoning in Farsi language models.

4 Experiments

This section details our evaluation protocol, the large-scale LLM baselines, the compact model and

fine-tuning recipe, and a quantitative and qualitative analysis of the results.

Data split. For all fine-tuning experiments we partition the **4,210** scenario–task bundles into **4,000** training items and a held-out set of **210** items for testing. Although the held-out set contains 210 scenarios, each yields multiple items, resulting in 1,470 evaluated items per model.

4.1 Evaluation metrics.

Each FarSense sub-task is scored with a metric that matches its answer format:

- **Categorical tasks** (*True/False* statements, single-blank *Cloze* answers and the **multiple-choice questions**) are evaluated by **accuracy**, i.e. exact match between the model output and the gold label.
- **Free-form generation tasks** (*Cause→Effect*, *Counterfactual*, *Explanation*, and the full-sentence *Cloze* completions) are scored with four complementary measures: BLEU-4 (Papineni et al., 2002), ROUGE-L F₁ (Lin, 2004), BERTScore F₁³ (Zhang et al., 2020), and cosine similarity⁴ (Reimers and Gurevych, 2019). Their arithmetic mean serves as the task score.

For each sub-task we compute a task-specific score as described above. To compare models at a higher level, we later aggregate these scores into track-level and overall scores, as detailed in the aggregation protocol below (§4.2). Cloze blanks typically expect a *single* token or short phrase, so exact string match is the most interpretable measure (treated as a single label). However, minor orthographic variants should not count as errors; therefore, we also log the text metrics above as a secondary figure.⁵

4.2 Large-scale LLM Results

This section reports the performance of the five flagship LLMs on the **210-scenario** FarSense test set. Each scenario includes two independent True/False statements (T/F-True and T/F-False), one 4-way multiple-choice question (MCQ), one cloze item (Cloze_cls), and four free-form generation items:

³Computed with `microsoft/deberta-xlarge-mnli`.

⁴Computed with sentence embeddings produced by HooshvarLab/bert-fa-base-uncased.

⁵This choice follows Yuan et al. (2021)’s recommendation for short answers.

Task	#Items	Avg. tok.	Median	Avg. TTR	Total tok.	Types	Corpus TTR
Cause–Effect	4,210	423	413	0.54	1.78 M	66 185	0.04
Cloze	4,210	426	416	0.53	1.79 M	65 899	0.04
Counterfactual	4,210	458	447	0.53	1.93 M	66 933	0.03
Explanation	4,210	453	443	0.53	1.91 M	66 886	0.04
Multiple-Choice	4,210	481	471	0.52	2.03 M	67 571	0.03
True/False 1	4,210	413	404	0.56	1.74 M	65 960	0.04
True/False 2	4,210	413	403	0.56	1.74 M	66 003	0.04
All tasks	29,470	438	429	0.54	12.91 M	71,347	0.01

Table 3: Token-level statistics for each FarSense subtask. Types counts unique lexical items after normalising Arabic and Farsi variants of the same character.

Metric	L3-1B		Mis-7B		L3-3B		L3-8B		Mis-N		Mis-24B	
	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
T/F-True Acc	0.767	0.905	0.743	0.919	0.476	0.929	0.843	0.929	0.924	0.957	0.962	0.981
T/F-True Prec	0.180	0.566	0.532	0.614	0.348	0.625	0.566	0.625	0.619	0.679	0.692	0.778
T/F-True Rec	0.261	0.756	0.771	0.959	0.488	0.963	0.920	0.963	0.961	0.978	0.980	0.990
T/F-True F1	0.169	0.590	0.489	0.664	0.239	0.681	0.573	0.681	0.672	0.752	0.768	0.852
T/F-False Acc	0.367	0.419	0.910	0.957	0.981	0.862	0.967	0.971	0.981	0.976	0.981	0.976
T/F-False Prec	0.229	0.257	0.810	0.881	0.948	0.814	0.881	0.889	0.912	0.914	0.930	0.916
T/F-False Rec	0.169	0.169	0.958	0.958	0.995	0.830	0.959	0.964	0.996	0.994	0.997	0.994
T/F-False F1	0.192	0.210	0.877	0.917	0.971	0.814	0.918	0.924	0.953	0.954	0.962	0.952
MCQ Acc	0.267	0.505	0.643	0.752	0.952	0.467	0.962	0.962	0.952	0.952	0.962	0.962
MCQ Prec	0.102	0.300	0.683	0.710	0.593	0.503	0.593	0.653	0.494	0.702	0.594	0.694
MCQ Rec	0.038	0.229	0.613	0.671	0.579	0.486	0.581	0.642	0.482	0.686	0.580	0.682
MCQ F1	0.059	0.259	0.577	0.690	0.586	0.477	0.586	0.647	0.488	0.694	0.587	0.687
Cloze Acc	0.052	0.067	0.333	0.544	0.543	0.368	0.576	0.697	0.586	0.737	0.543	0.667
Cloze Prec	0.027	0.036	0.201	0.392	0.373	0.225	0.406	0.481	0.416	0.543	0.373	0.462
Cloze Rec	0.027	0.036	0.201	0.392	0.375	0.225	0.406	0.481	0.416	0.543	0.375	0.462
Cloze F1	0.027	0.036	0.201	0.392	0.374	0.225	0.406	0.481	0.416	0.543	0.374	0.462

Table 4: Fine-tuning impact on classification-style tasks.

Cause→Effect, Counterfactual, Explanation, and Cloze_text.

Fine-grained Classification Metrics

- **GPT-4o** effectively dominates the binary statements and MCQs, achieving the strongest overall precision–recall–F1, while **DeepSeek-Chat** slightly outperforms it on some T/F precision metrics and attains the best CLOZE_{CLS} accuracy.
- **DeepSeek-Chat** remains unrivalled on short CLOZE_{CLS} confirming its strong grounding of lexical knowledge.
- All models lose ∼40 pp when moving from MCQ to Cloze, highlighting the difficulty of pinpointing the exact Farsi token.

Free-form Generation Metrics Free-form reasoning is orders of magnitude harder. Even the best BLEU for CAUSE→EFFECT is only 0.11, signalling that causal commonsense in Farsi remains a major research frontier. DeepSeek-Chat and Gemini 2.5-Flash share the top spot, each surpassing GPT-4o by ≈ 2 pp in the macro average of the four

similarity metrics.

Aggregation protocol. For each model we report three summary scores. The **Classification** score is the macro-average of the accuracies on the four discrete tasks (T/F-True, T/F-False, MCQ, Cloze_{cls}) from Table 7. The **Free-form** score is the macro-average of the four similarity means on Cause→Effect, Counterfactual, Explanation, and Cloze_{text} from Table 8. The **Overall** score in Table 9 is then the arithmetic mean of these two track-level scores:

$$\text{Overall} = \frac{1}{2}(\text{Classification} + \text{Free-form}).$$

- **DeepSeek-Chat** retains first place with an overall 0.700, edging Gemini-2.5-Flash by 0.004 pp. Its gains stem from Cloze precision and a slight edge on causal generation.
- **Gemini-2.5-Flash** matches DeepSeek on free-form reasoning, confirming the robustness observed during dataset construction.
- **GPT-4o** excels at MCQ and binary facts but is

Metric	L3-1B		Mis-7B		L3-3B		L3-8B		Mis-N		Mis-24B	
	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
Cause–Effect BLEU	0.005	0.014	0.030	0.071	0.014	0.026	0.052	0.065	0.033	0.060	0.041	0.090
Cause–Effect ROUGE	0.042	0.126	0.176	0.310	0.112	0.177	0.235	0.293	0.199	0.300	0.222	0.343
Cause–Effect BERT	0.095	0.492	0.553	0.624	0.267	0.518	0.583	0.623	0.547	0.626	0.587	0.643
Cause–Effect Cos	0.548	0.641	0.700	0.771	0.629	0.690	0.713	0.770	0.721	0.773	0.728	0.793
Counterfactual BLEU	0.005	0.014	0.021	0.054	0.019	0.026	0.027	0.050	0.024	0.063	0.028	0.064
Counterfactual ROUGE	0.039	0.127	0.159	0.280	0.097	0.160	0.192	0.241	0.153	0.258	0.185	0.291
Counterfactual BERT	0.089	0.483	0.545	0.620	0.258	0.506	0.573	0.616	0.536	0.623	0.575	0.639
Counterfactual Cos	0.662	0.780	0.815	0.874	0.736	0.823	0.837	0.870	0.842	0.890	0.851	0.907
Explanation BLEU	0.004	0.017	0.021	0.046	0.019	0.029	0.026	0.045	0.022	0.053	0.028	0.054
Explanation ROUGE	0.039	0.152	0.157	0.299	0.096	0.167	0.187	0.250	0.150	0.261	0.183	0.295
Explanation BERT	0.090	0.486	0.543	0.622	0.258	0.509	0.580	0.623	0.538	0.632	0.574	0.642
Explanation Cos	0.644	0.796	0.814	0.885	0.723	0.837	0.847	0.888	0.853	0.910	0.863	0.919
Cloze-text BLEU	0.114	0.183	0.298	0.330	0.334	0.358	0.394	0.411	0.403	0.453	0.370	0.411
Cloze-text ROUGE	0.440	0.533	0.638	0.720	0.611	0.638	0.675	0.699	0.675	0.740	0.605	0.675
Cloze-text BERT	0.699	0.756	0.786	0.819	0.756	0.780	0.796	0.811	0.819	0.846	0.768	0.819
Cloze-text Cos	0.798	0.836	0.846	0.873	0.857	0.871	0.870	0.885	0.893	0.907	0.857	0.893

Table 5: Fine-tuning impact on free-form generation tasks.

Metric	L3-1B		Mis-7B		L3-3B		L3-8B		Mis-N		Mis-24B	
	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT	Base	FT
Classification	0.286	0.582	0.615	0.817	0.593	0.732	0.761	0.813	0.789	0.826	0.810	0.845
Free-form	0.168	0.319	0.377	0.482	0.305	0.326	0.435	0.471	0.419	0.482	0.445	0.499
Overall	0.227	0.450	0.496	0.649	0.449	0.529	0.598	0.642	0.604	0.654	0.627	0.672

Table 6: Aggregated performance before and after fine-tuning.

penalised by its lower Cloze accuracy, pointing to tokenisation issues with undiacritised Farsi.

- A perfect commonsense model is still far away: even the leader covers just 70 % of the combined task ceiling, leaving ample room for future Farsi-centric research.

4.3 Fine-tuning on FarSense

To quantify how much task-specific supervision can improve each open model relative to its own zero-shot baseline and approach flagship scores, we fine-tuned six open-weights, instruction-tuned backbones—Llama-3.2 (1B)(Grattafiori et al., 2024), Llama-3.2 (3B), Llama-3.1 (8B), Mistral (7B)(Jiang et al., 2023), Mistral-Nemo (12B), and Mistral-Small (24B)—spanning the 1 to 24 billion parameter range, on the train split of **FarSense** (4,000 scenario–task bundles; see Section 4). Details of the experiment, including the training and generation hyperparameters, are provided in Appendix G.

Tables 4–6 contrast *Base* vs. *FT* across classification metrics, free-form generation metrics, and the macro aggregates from. Model names are shortened to L3 (LLAMA 3.X) and Mis (MISTRAL) for brevity.

Fine-tuning yields consistent improvements for ev-

ery model: the mean absolute jump is $\approx +10$ pp on the overall score. The smallest model, **L3-1B**, benefits most (+23 pp), followed by **Mis-7B** (+16 pp). Even the strongest baseline, **Mis-24B**, still gains about +5 pp.

Source of improvement. Gains are largest in classification ($\approx +13$ pp on average), driven by sharper True/False precision and Cloze accuracy—evidence that direct exposure to **FarSense** answer formats helps probability calibration. Free-form reasoning also improves ($\approx +7$ pp), though absolute scores on *Cause → Effect* and *Counterfactual* remain low, confirming that causal commonsense in Farsi is still an open challenge.

5 Conclusion & Future Work

A human-in-the-loop LLM pipeline can **reliably generate, vet, and use** a large-scale Farsi commonsense dataset. Combining prompt-based generation with two native-speaker passes yielded **FarSense**, a 29k-item benchmark that both *evaluates* and *improves* Farsi models. For **RQ1**, Gemini-2.5-Flash, GPT-4o, and Mistral-Large produced the strongest outputs; Gemini-2.5-Flash led (2.6/3 for scenarios, 2.3/3 for tasks), though 78% of bundles still required edits, confirming the need for human oversight. In **RQ2**, a two-stage validation

Model	T/F-TRUE				T/F-FALSE				MCQ				CLOZE _{CLS}			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Mistral-Large	0.967	0.765	0.990	0.854	0.990	0.816	0.988	0.890	0.957	0.592	0.563	0.575	0.543	0.373	0.375	0.374
Gemini-2.5-Flash	0.976	0.790	0.992	0.866	0.990	0.838	0.990	0.908	0.952	0.594	0.579	0.586	0.576	0.406	0.406	0.406
DeepSeek-Chat	0.981	0.833	0.993	0.889	0.995	0.885	0.996	0.937	0.962	0.593	0.581	0.586	0.586	0.416	0.416	0.416
GPT-4o	0.986	0.832	0.995	0.907	0.995	0.901	0.996	0.947	0.967	0.741	0.727	0.734	0.490	0.326	0.326	0.326
Qwen-Plus	0.986	0.806	0.993	0.888	0.990	0.864	0.992	0.923	0.952	0.494	0.482	0.488	0.452	0.292	0.292	0.292

Table 7: Classification-style performance (\uparrow) on the four discrete FarSense sub-tasks. Best numbers per column are **bold**.

Model	Cause→Effect				Counter-factual				Explanation				Cloze _{text}			
	BLEU	ROUGE	BERT	Cos	BLEU	ROUGE	BERT	Cos	BLEU	ROUGE	BERT	Cos	BLEU	ROUGE	BERT	Cos
Mistral-Large	0.092	0.307	0.627	0.773	0.047	0.246	0.597	0.817	0.044	0.250	0.598	0.809	0.370	0.605	0.768	0.857
Gemini-2.5-Flash	0.110	0.357	0.645	0.795	0.064	0.269	0.615	0.846	0.065	0.287	0.615	0.862	0.394	0.637	0.774	0.870
DeepSeek-Chat	0.110	0.340	0.645	0.795	0.064	0.262	0.615	0.846	0.065	0.282	0.615	0.864	0.403	0.675	0.819	0.893
GPT-4o	0.093	0.318	0.632	0.782	0.053	0.253	0.604	0.839	0.050	0.258	0.604	0.852	0.346	0.620	0.771	0.844
Qwen-Plus	0.079	0.271	0.591	0.769	0.041	0.245	0.596	0.806	0.041	0.251	0.590	0.823	0.370	0.605	0.768	0.857

Table 8: Free-form generation metrics (\uparrow). Each sub-task score is the mean of the four columns; these means are used in §4.2.

Model	Classification	Free-form	Overall
DeepSeek-Chat	0.881	0.518	0.700
Gemini-2.5-Flash	0.874	0.518	0.696
Mistral-Large	0.864	0.497	0.681
GPT-4o	0.860	0.494	0.677
Qwen-Plus	0.845	0.481	0.663

Table 9: Headline macro scores (\uparrow). Classification is the average of the four accuracies in Table 7; Free-form is the average of the four means in Table 8.

(edit + consensus) corrected 60.1% and regenerated 17.9%, achieving high precision at a fraction of manual cost and proving scalability to other low-resource languages. For **RQ3**, models reach near-ceiling accuracy (≈ 0.95 –0.97) on binary and multiple-choice items, but even DeepSeek-Chat (0.70 macro) struggles on causal, counterfactual, and *Cloze_text* tasks ($\text{BLEU} \leq 0.11$). Finally, **RQ4** shows that fine-tuning six compact models on FarSense yields consistent gains—especially for the smallest backbones—demonstrating its value as a training signal for resource-constrained Farsi models, though a performance gap to flagship LLMs persists. **Next Steps.** We will extend FarSense to Dari, Tajiki, and other regional dialects, expand it to multimodal (image / audio-grounded) reasoning, and release human + model rationales for chain-of-thought evaluation. Annual re-benchmarking and full release of data, scripts, and checkpoints will promote reproducible, resource-efficient commonsense research across under-represented languages.

6 Limitations

In line with ACL policy, we outline the principal shortcomings of our work. We refrain from present-

ing these points as future work or adding citations, and we avoid emphasising the strengths of our approach.

Domain and topic bias. All scenarios originate from lead paragraphs in Farsi Wikipedia. Although we sampled from “featured” and “vital” article indices, Wikipedia skews toward literate, urban and male-centred topics. Consequently, FarSense may under-represent everyday experiences of rural communities, children, and marginalised social groups.

Dialectal coverage. Our prompting strategy instructed the language models to generate responses in standard Farsi, without explicit coverage of dialectal variants such as Dari, Tajiki, Luri, or other regional Iranian varieties. As a result, model performance on FarSense may overestimate general capabilities, particularly when applied to dialectal or non-standard inputs.

Synthetic-data artefacts. Each item was first machine-generated, then post-edited by two annotators. Subtle factual inaccuracies, unnatural phrasings or culturally inappropriate assumptions can still persist—especially in causal or counterfactual prompts where human validation is harder. Users of FarSense should manually inspect samples before drawing fine-grained linguistic conclusions.

Annotation scale and diversity. Only two native speakers performed all validations. Their demographic homogeneity (age, education level, home dialect) limits the cultural breadth of error detection and may allow implicit biases—e.g. gender stereotypes—to remain.

Metric limitations. For free-form answers we average BLEU, ROUGE, BERTScore and embedding cosine similarity. These surface-level metrics correlate imperfectly with human judgments of commonsense plausibility; high scores do not guarantee deep reasoning, while low scores may penalise legitimate paraphrases.

Environmental cost. All computational stages—large-scale scenario and task generation as well as the three-epoch fine-tuning of six compact models—were run exclusively on NVIDIA RTX A6000 GPUs. Generation consumed about 75 GPU-hours, and fine-tuning added roughly 108 GPU-hours (35.8 wall-clock hours on a 3-GPU A6000 node). While modest by today’s large-model standards, this compute budget may still be prohibitive for some research groups and carries a non-negligible carbon footprint.

Potential for misuse. FarSense could be mis-applied as a high-stakes diagnostic of human traits (e.g. educational testing) even though it was never validated for such scenarios. The benchmark should be used solely for research on machine commonsense reasoning in Farsi.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, and 1 others. 2023. [Qwen technical report](#). ArXiv preprint arXiv:2309.16609.

Ted Byrt, Jeremy Bishop, and John B. Carlin. 1993. [Bias, prevalence and kappa](#). *Journal of Clinical Epidemiology*, 46(5):423–429.

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, and 1 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). ArXiv preprint arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bingxuan Wang, and 1 others. 2025. [Deepseek-v3 technical report](#). ArXiv preprint arXiv:2412.19437.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#). In *NAACL*.

Omrid Ghahroodi, Marzia Nouri, Mohammad V. Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad H. Rohban. 2024. [Khayyam challenge \(persianmmlu\): Is your llm truly wise to the persian language?](#) ArXiv preprint arXiv:2309.xxxxx.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. [The llama 3 herd of models](#). ArXiv preprint arXiv:2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and 1 others. 2023. [Mistral 7b](#). ArXiv preprint arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension](#). In *ACL*.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, and 1 others. 2021. [Parsinlu: A suite of language understanding challenges for persian](#). *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, and et al. 2018. [The narrativeqa reading comprehension challenge](#). In *TACL*.

Salima Lamsiyah, Kamyar Zeinalipour, Matthias Brust, Marco Maggini, Pascal Bouvry, Christoph Schommer, and 1 others. 2025. [ArabicSense: A benchmark for evaluating commonsense reasoning in arabic with large language models](#). In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 1–11.

Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81. ACL.

Mistral AI. 2025. [Mistral-large](#). <https://mistral.ai/news/mistral-large>. Accessed: 2025-11-09.

OpenAI, Josh Achiam, Sandhini Agarwal, Lama Ahmad, and 1 others. 2024. [Gpt-4 technical report](#). ArXiv preprint arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal commonsense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [Squad: 100,000+ questions for machine comprehension of text](#). In *EMNLP*.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992. ACL.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Hamidreza Saffari, Mohammadamin Shafiei, and Francesco Pierri. 2024. PSN: Persian social norms dataset for cross-cultural AI. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024a. **mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans**. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024b. Multilingual commonsenseqa (mcsqa) with unified creation strategy by language models and humans. *arXiv preprint arXiv:2406.04215*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. **Social IQa: Commonsense reasoning about social interactions**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4463–4473.

Mehrnoosh Shamsfard. 2019. Challenges and opportunities in processing low resource languages: A study on persian. In *Proc. of the Int. Conf. on Language Technologies for All (LT4All)*.

Mehrnoosh Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, Seyed M. H. Hashemi, and 1 others. 2024. Farseval-pkbets: A new diverse benchmark for evaluating persian large language models. ArXiv preprint arXiv:2504.14690.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. **CommonsenseQA: A question answering challenge targeting commonsense knowledge**. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 4149–4158.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. **Semeval-2020 task 4: Commonsense validation and explanation**. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval 2020)*, pages 307–321.

Wikipedia contributors. 2025a. Wikipedia: 100 vital articles. *Wikipedia (fa) – 100 Vital Articles*. Accessed: 2025-11-09.

Wikipedia contributors. 2025b. Wikipedia: Featured articles. *Wikipedia (fa) – Featured Articles*. Accessed: 2025-11-09.

Wikipedia contributors. 2025c. Wikipedia: List of articles every wikipedia should have. *Wikipedia (fa) – Articles Every Wikipedia Should Have*. Accessed: 2025-11-09.

Wikipedia contributors. 2025d. Wikipedia: List of most-viewed articles by topic. *Wikipedia (fa) – Most-Viewed by Topic*. Accessed: 2025-11-09.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. <https://github.com/neulab/BARTScore>. NeurIPS 2021 Poster. Last modified: 26 May 2025.

Kamyar Zeinalipour, Achille Fusco, Asya Zanollo, Marco Maggini, and Marco Gori. 2024a. Harnessing LLMs for educational content-driven italian crossword generation. *arXiv preprint arXiv:2411.16936*.

Kamyar Zeinalipour, Neda Jamshidi, Fahimeh Akbari, Marco Maggini, Monica Bianchini, Marco Gori, and 1 others. 2025a. 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. Association for Computational Linguistics.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, and Marco Gori. 2024b. Automating turkish educational quiz generation using large language models. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 246–260. Springer.

Kamyar Zeinalipour, Yusuf Gökberk Keptiğ, Marco Maggini, Leonardo Rigutini, and Marco Gori. 2024c. A turkish educational crossword puzzle generator. In *International Conference on Artificial Intelligence in Education*, pages 226–233. Springer.

Kamyar Zeinalipour, Mehak Mehak, Fatemeh Parsamotamed, Marco Maggini, and Marco Gori. 2024d. Advancing student writing through automated syntax feedback. In *International Workshop on AI in Education and Educational Research*, pages 52–66. Springer.

Kamyar Zeinalipour, Moahmmad Saad, Marco Maggini, and Marco Gori. 2025b. **From Arabic text to puzzles: LLM-driven development of Arabic educational crosswords**. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 479–495, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kamyar Zeinalipour, Mohamed Zaky Saad, Oumaima Attafi, Marco Maggini, and Marco Gori. 2025c. Shawarma chats: A benchmark exact dialogue & evaluation platter in egyptian, maghrebi & modern standard arabic—a triple-dialect feast for hungry language models. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 472–524.

Kamyar Zeinalipour, Mohamed Zaky Saad, Marco Maggini, and Marco Gori. 2025d. From arabic text to puzzles: Llm-driven development of arabic educational crosswords. *arXiv preprint arXiv:2501.11035*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. **NormBank: A knowledge bank of situational social norms**. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 7756–7776.

Andrea Zugarini, Kamyar Zeinalipour, Achille Fusco, and Asya Zanollo. 2024a. Ecwca-educational crossword clues answering: A calamita challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1239–1244.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024b. Clue-instruct: Text-based clue generation for educational crossword puzzles. *arXiv preprint arXiv:2404.06186*.

A Scenario and Tasks in Full Detail

The following pages present the complete texts of the scenarios and subtasks referenced in Figures 2 and 1, which previously showed only brief excerpts. This includes the original Wikipedia passage, the generated scenario, and all six associated tasks.

B Prompts (full text)

```
You will receive one short Farsi passage.

1. Extract every explicit factual statement from the passage (keep them concise and literal no interpretation).
2. Convert each fact into a plausible, everyday commonsense scenario that implies the fact without stating it directly.
3. Using every fact from step 1, expand the scenario from step 2 into a 100 to 300 word Farsi narrative, adding natural detail or a minor hypothetical twist. Be sure every extracted fact is clearly included.

Return only the narrative from step 3 no labels, explanations, JSON, or extra text of any kind.

Text:
{{WikiPedia_Paragraph}}
```

Listing 1: Scenario Re-writing Prompt

```
For the following narrative, create all six commonsense reasoning task types below, entirely in Farsi, each with its correct answer:
Make sure the question requires implicit understanding and is not answerable through rote memorization.

1) Multiple-choice QA
    Write one commonsense question with four choices (A - D)
    The question should require reasoning beyond direct sentence recall.
    Provide the correct option letter.

2) True/False statements
    Write two declarative statements:
    One must be clearly true, based on commonsense reasoning.
    One must be clearly false, based on contradiction or implausibility.
    Each item: {"statement": "...", "answer": true|false}.

3) Cause-and-Effect pairs
    - Write one plausible cause-effect pair based on the story.
    - Focus on why something happened or what its result was.
    Use commonsense logic, not just temporal order.
    One pair: {"cause": "...", "effect": "..."}.

4) Cloze completion
    Choose a meaningful sentence and blank out a key word or phrase using <---->.
    The answer must be non-obvious without understanding the context.
    Provide the missing text as "answer".

5) Explanation generation
    Ask a Why or How question about a behavior, intention, or event.
    Answer it in 1-2 short Farsi sentences.
    The answer should show inference, not surface repetition.

6) Counterfactual reasoning
    Ask a "What_if_..." question that changes one detail in the scenario.
    Predict a reasonable alternate outcome.
```

Wikipedia text

گریستن یکی از واکنش‌های رفتاری است. باید توجه شود که گریستن با تولید اشک متفاوت است. اشک و واکنش چشم به یک حرکت شیمیایی مانند بوی پیاز یا گاز اشک‌آور است و مواد شیمیایی موجود در اشک و گریه با هم تفاوت دارند؛ مانند اشک که با ییدن یک فیلم غم‌انگیز پیده می‌آید از اشک حاصل از بوی پیاز، پروتئین پیشتری دارد.

گریه با خاری شدن اشک از چشم و در سیاری موارد بالله و تغیر حالت دهان همراه است. گریه و واکنش و نمود ایجاد وضعیت احساسی است. پیشتر گریه در موقعی که فرد غمگین است صورت می‌گیرد.

معمولًا گریه‌ای که در اثر ناراحتی پیده می‌آیند حاصل این هستند که مغز یا فلسفه ذهن نمی‌تواند توجیه و توصیفی برای مشکل یا ناراحتی به وجود آمده پیدا کند و برای تخلیه احساسات گریه می‌کند. از این رو گریه‌ای آدم هایی که مرتب گریه می‌کنند افسرده هستند، مخصوص این است که مغز اغفار نمی‌تواند هیچ توجیهی برای مشکلات پیدا کند. گریه، حالتی است که از برخورد انسان با یک سلسله واقعیت‌ها پیدا می‌آید و ریشه در سوز دل و شکستن قلب یا شادی روزگارون دارد.

Crying is a behavioral response. It should be noted that crying is different from tear production. Tears are the eyes' reaction to a chemical stimulus such as the smell of onions or tear gas, and the chemical composition of emotional tears differs from reflex tears; for example, tears shed while watching a sad movie contain more protein than those caused by cutting onions. Crying involves the flow of tears from the eyes, and often includes sobbing and changes in facial expression. It is a reaction that reflects an emotional state, typically occurring when a person feels sadness. Crying due to emotional distress usually happens when the brain or mind fails to find an explanation or justification for a troubling situation, leading to crying as a way to release emotions. This is why people who cry frequently or are depressed often do so because their minds are overwhelmed and unable to rationalize their problems. Crying emerges from confronting certain realities and stems from heartbreak, deep sorrow, or even overwhelming joy.

Scenario

سara دیشب فیلم غم‌انگیزی دید و بی‌اختیار اشک از چشمانش جاری شد. او که معمولاً فقط وقتی واقعاً غمگین است گریه می‌کند، متوجه شد که این یک واکنش رفتاری است. اینهاش می‌لرزید و نالمهای ارامی از گلویش بیرون می‌آمد. این شنایهای واضح از وضعیت احساسی او بود. سارا به باد اورد روزی را که پیاز خرد می‌کرد و چشمانش بی‌دلیل اشک‌الود شده بود. او می‌دانست که گریستن با صرف تولید اشک فرق ندارد. اشک‌های پیاز، فقط واکنش شیمیایی بود و هیچ احساسی پشت آن نبود. به گفته استاد زیست‌شناسی، اشک فیلم غمگین پروتئین پیشتری نسبت به اشک پیاز داشت. سارا فکر کرد، انسان در مواجهه با واقعیت‌ها می‌گرید. زمانی که مغزش نتوانست برای مشکل پیش آمده توجیهی پیدا کند، گرفست تا احساساتش تخلیه شود. یاد دوستش افتاد که مدام گریه می‌کرد؛ شاید مغز او انقدر به بی‌پیش رسانیده بود که دیگر نمی‌توانست هیچ توجیهی برای مشکلاتش بیابد. در نهایت، سارا دریافت که ریشه گریه، چه از دلشکستگی باشد و چه از شادی بی‌حد و حصر، همیشه عمیق است.

Last night, Sara watched a sad movie and tears flowed from her eyes uncontrollably. She, who usually cries only when she is truly sad, realized that this was a behavioral reaction. Her lips were trembling, and soft sobs came from her throat—clear signs of her emotional state. She recalled a day when she was chopping onions and her eyes had teared up for no reason. She knew that crying is different from simply producing tears. Onion-induced tears were just a chemical reaction, with no emotion behind them. According to her biology professor, tears shed during a sad movie contain more protein than onion tears. Sara thought about how humans cry when faced with reality. When the brain fails to find a justification for a problem, it cries to release emotions. She remembered her friend who cried constantly; perhaps his mind had hit such a dead end that it could no longer explain his struggles. In the end, Sara realized that the root of crying—whether from heartbreak or overwhelming joy—is always deep.

Task 1: Multiple Choice Question

سارا از چه نظر گریه ناشی از تماشای فیلم غم انگیز را با گریه ناشی از خرد کردن پیاز متفاوت می‌داند؟
Q: In what way does Sara consider crying from watching a sad movie different from crying while chopping onions?

A) میزان پروتئین اشک‌ها

A: The protein content of the tears

B) وجود یا عدم وجود احساسات عمیق پشت آن

B : The presence or absence of deep emotions behind it

C) شدت لرزش لب‌ها

C: The intensity of lip trembling

D) توانایی مغز برای پافتن توجیه

D : The brain's ability to find justification

Correct Answer: B) وجود یا عدم وجود احساسات عمیق پشت آن

Correct Answer B: The presence or absence of deep emotions behind it

Task 2: True/False judgment

اشک‌هایی که سارا هنگام خرد کردن پیاز ریخت، نشانه‌ای از ناراحتی عاطفی او نبود.

Statement : The tears Sara shed while chopping onions were not a sign of emotional distress.

Answer: True

گریه کردن فقط در مواجهه با اتفاقات ناخوشایند رخ می‌دهد.

Statement : Crying only happens in response to unpleasant events.

Answer: False

Task 3: Cause -Effect inference

مغز انسان نتواند برای مشکل پیش آمده توجیهی بیندا کند

Cause : The human brain cannot find a justification for a problem that has occurred

احساسات فرد تخلیه شود

Effect : The person's emotions are released

Task 4: Completion

ashk-hayi ke hen-gam xrd krdn biyaz tolid mi-shod, chrafa-ye pak va kntsh shimiayi ast. Sentence: The tears produced while chopping onions are merely a chemical reaction and lack "----".

Answer: احساس

Answer : emotion

Task 5: Explanation

Q: چرا سارا فکر می کند دوستش که مدام گریه می کند، ممکن است با مشکلات عمیق تری روبرو باشد؟
Q : Why does Sara think that her friend who cries constantly might be facing deeper problems?

A: سارا معتقد است گریه زمانی اتفاق می افتد که مغز نتواند برای مشکلات راه حلی پیدا کند. گریه مداوم دوستش نشانه ای از ناتوانی مغز او در یافتن توجیه یا راهکار برای مشکلات است.

A : Sara believes that crying happens when the brain cannot find a solution to problems. Her friend's constant crying is a sign that his brain is unable to justify or cope with his issues.

Task 6: Counterfactual reasoning

Q: اگر سارا به جای فیلم غم انگیز، یک کمدی بسیار خنده دار تماشا کرده بود، آیا ممکن بود باز هم اشک بریزد؟
Q : If Sara had watched a very funny comedy instead of a sad movie, could she still have cried?

A: بله، بسیار ممکن بود. سارا در پایان متن اشاره می کند که گریه می تواند ناشی از شادی بی حد و حصر نیز باشد و ریشه های عمیقی دارد.

A : Yes, it is quite possible. Sara mentions at the end that crying can also result from overwhelming joy and has deep roots.

Quality gates **apply** to every item: the correct answer must never appear verbatim **in** the narrative (the no-span rule); each item must include a key named 'cs_skill' whose value **is** one of physical, social, temporal, intentional, causal, **or** deontic to state which commonsense capacity **is** being exercised; each item must also contain a single-sentence Farsi explanation under 'why_cs' that clarifies what piece of world knowledge **is** required. If **any** item fails one of these three requirements, delete it **and** regenerate until it conforms. A weak example would be asking the hero's age when it is explicitly given in the text; a strong example would ask why the hero chose the more distant bakery, which requires understanding a social norm about quality preferences.

 OUTPUT_FORMAT--RETURN_EXACTLY_THIS_JSON

```
{
  "multiple_choice": {
    "question": "...",
    "options": ["A) ...", "B) ...", "C) ...", "D) ..."],
    "answer": "B",
    "cs_skill": "social",
    "why_cs": "..."
  },
  "true_false": [
    {"statement": "...", "answer": true, "cs_skill": "social", "why_cs": "..."},
    {"statement": "...", "answer": false, "cs_skill": "social", "why_cs": "..."}
  ],
  "cause_effect": [
    {"cause": "...", "effect": "...", "cs_skill": "social", "why_cs": "..."}
  ],
  "cloze": [
    {"sentence": "...<____>...", "answer": "...", "cs_skill": "social", "why_cs": "..."}
  ],
  "explanation": {
    "prompt": "...",
    "answer": "...",
    "cs_skill": "social",
    "why_cs": ...
  },
  "counterfactual": {
    "question": "...?",
    "answer": "...",
    "cs_skill": "social",
    "why_cs": ...
  }
}

Do_not_output_anything_outside_this_JSON_object.
The_JSON_must_be_valid_and_parseable.

Text:
{{Senario}}}
```

Listing 2: Task Derivation Prompt

C Human Evaluation Guidelines

Each item (scenario or task) should be rated as **A**, **B**, or **C**, based on its quality, clarity, and alignment with the original Farsi content. After evaluating each item, annotators must assign:

- A final rating (**A**, **B**, or **C**).

Scenario Text Evaluation

- **Grade A:** A well-formed, coherent narrative that captures most of the key information from the original Wikipedia article. It reads naturally, is grammatically sound, and avoids English words or unnatural expressions.

- **Grade B:**

The output is generally coherent, informative, and aligned with the topic, but it has some shortcomings that prevent it from achieving top quality. These may include:

- Slightly awkward phrasing or minor grammatical issues
- Soft or unclear metaphors that reduce clarity
- A story that is topically aligned but omits some important details from the original text
- A text that contains all the relevant details but lacks a natural story-like flow and instead resembles a descriptive passage or question-answer format

- **Grade C:** The structure is weak, content is inaccurate or off-topic, or the tone is overly exaggerated or unrealistic. Also applies when the text is missing, in English, or uses hard-to-follow metaphors.

Multiple-Choice Questions

- **Grade A:** Properly formatted (1 question + 4 options), logically consistent, and clearly grounded in the scenario or commonsense. Fluent and grammatical.
- **Grade B:** Follows the correct format and remains answerable, but the phrasing is slightly awkward, or the logic is less precise. Generally clear and on-topic.
- **Grade C:** Disorganized, unclear, or not connected to the scenario. The language may be confusing, overly metaphorical, or missing altogether.

True/False Statements

- **Grade A:** A concise and accurate claim with a clear truth value, well-aligned with the original content or commonsense.
- **Grade B:** The statement is still valid and clearly related, but phrasing may be slightly imprecise or contain mild metaphorical distractions. Overall understandable.
- **Grade C:** The statement is vague, irrelevant, unrealistic, or structurally broken. Also applies if left blank or poorly generated or not generated.

Cloze (Fill-in-the-Blank) Questions

- **Grade A:** A fluent and complete sentence with a meaningful blank that is clearly inferable from the text or commonsense.
- **Grade B:** Mostly understandable and contextually appropriate, but phrasing may feel slightly forced or less polished. Still logically constructed.
- **Grade C:** Incoherent, unrelated, confusing, or not generated. May also include awkward metaphors or illogical sentence structure or not generated.

Cause-Effect Questions

- **Grade A:** A natural and logical cause-effect relation clearly grounded in the scenario or general knowledge.
- **Grade B:** The causal relationship is generally correct and relevant, but phrasing may lack fluency or contain small inconsistencies. Still logical.
- **Grade C:** Illogical or confusing relation, weak connection to the source, or severely flawed wording. Also includes empty or non-generated output.

Explanation Questions

- **Grade A:** Directly asks for a reason, detail, or concept from the text; clearly worded and well-targeted.
- **Grade B:** Still addresses a relevant idea and remains understandable, but may suffer from slight vagueness or awkward formulation.
- **Grade C:** Unclear purpose, poor language, or unrelated to the scenario. Also applies if missing or dominated by confusing metaphors.

Counterfactual Questions

- **Grade A:** A plausible and meaningful hypothetical based on the scenario (e.g., “What if...”), with fluent structure and clear logic.
- **Grade B:** The counterfactual is still reasonable and text-related, but may be phrased less smoothly or include minor flaws. Understandable overall.
- **Grade C:** Implausible, unclear, disconnected, or exaggerated. Also applies if not generated or if difficult metaphors interfere with clarity.

D Full Dialogue Examples

Below, we present representative examples for each models: GPT-4o, Gemini-2.5-Flash, Mistral-Large, Qwen-Plus, and DeepSeek-Chat. Each example includes the *original Wikipedia passage*, the *generated scenario* derived from it, and a set of *six tasks* constructed from the scenario. All items were evaluated by human annotators and assigned a quality rating of **A**, **B**, or **C**, based on the criteria described in our evaluation guidelines.

Chatgpt:

دریاچه آرال (به انگلیسی: Aral Lake) یا دریاچه خوارزم (به انگلیسی: Kharazm Lake) یک دریاچه آب شور در آسیای میانه، میان قرقاولیاستان در کشور ازبکستان در جنوب و قراقستان در شمال بود. دریای آرال میلیون‌ها سال پیش بخشی از دریای پارانتس بود. دریای آرال پک دریاچه اندره هیک (پرشور) بود که بین قراقستان در شمال و ازبکستان در جنوب آن قرار داشت که در دهه ۱۹۶۰ شروع به کوچک شدن کرد و در دهه ۲۰۱۰ تا حد زیادی خشک شد. در منطقه آکتوپه و کیزیلوردا قراقستان و منطقه خودمختار کاراکالپاکستان ازبکستان قرار داشت. این نام تقریباً از زبان‌های مغولی و ترکی به «دریای چز ایر» ترجیح می‌شود، که اشاره‌ای به تعداد زیادی جزیره (پیش از ۱۱۰۰) است که زمانی در آب‌های آن پراکنده بودند. حوضه آبریز دریای آرال ازبکستان و بخش‌هایی از افغانستان، ایران، قراقستان، قرقیزستان و ترکمنستان را در بر می‌گیرد. دریاچه آرال که قبلاً چهارمین دریاچه بزرگ جهان با مساحت ۶۸۰۰ کیلومتر مربع (۲۶۰۰ مایل مربع) بود، در دهه ۱۹۶۰ پس از منحرف شدن رودخانه‌ایی که از آن تغذیه می‌کردند توسط پروژه‌های آبیاری شور و شروع به کوچک شدن کرد. تا سال ۲۰۰۷، به ۱۰٪ از اندازه اصلی خود کاهش یافت و به چهار کهارمند دریاچه تغییم شد: دریای آرال شمالی، حوضه‌های شرقی و غربی دریای آرال جنوبی که زمانی بسیار بزرگتر بود، و دریاچه میانی کوچکتر بود. دریاچه نازک در لبه غربی دریای جنوبی ساقی عقب‌نشینی کرد. در سال‌های پس از آن، جزیان‌های گامیه‌گاه آب سبب تجدید دریاچه جنوب شرقی شده است که گاهی به میزان کمی دوباره پر می‌شود. تصاویر ماهواره‌ای ناسا در اوایل ۲۰۱۴ نشان داد که برای اولین بار در تاریخ مدرن حوضه شرقی دریای آرال کاملاً خشک شده است. حوضه شرقی اکنون صحرای آرالک نامیده می‌شود. در تلاشی در قراقستان برای نجات و دوباره پر کردن دریای آرال شمالی، سدایک کوکارال در سال ۲۰۰۵ تکمیل شد. در سال ۲۰۰۸ سطح آب ۱۲ متر (۴۰ فوت) بالاتر از سال ۲۰۰۳ بود. از سال ۲۰۱۳، شوری کاهش یافت و سامه‌ها دوباره به تعداد کافی برخی از ماهیگیری‌ها حضور داشتند. پس از بازدید از مویناک در سال ۲۰۱۱، بن کی مون، بیرونی پیشین سازمان ملل، کوچک شدن دریای آرال را «یکی از بزرگ‌ترین بلایی زیست‌محیطی سیار» نامید. صنعت ماهیگیری منطقه که زمانی پر رونق بود، ویران شده و بیکاری و مشکلات اقتصادی را به همراه داشته است. آب رودخانه سیر دریا برای آبیاری حدود میلیون هکتار (۵۰۰۰ هکتار) از زمین‌های کشاورزی در دره فرغانه استفاده می‌شود. منطقه دریای آرال به شدت آلوده است و به دنبال آن مشکلات بهداشت عمومی جدی است. پیشکو اسناد تاریخی مربوط به دریای آرال را به عنوان منبعی برای مطالعه تراژی زیست‌محیطی به فهرست حافظه جهانی خود اضافه کرده است.

در گذشته‌های دور، یک دریاچه پیشوار با آب‌های شور در قلب آسیای میانه وجود داشت که مردمان محلی به آن «دریای چز ایر» می‌گفتند. این نام به خاطر هزاران جزیره کوچک و بزرگی بود که مانند مرواریدهایی در پهنه آبی آن پراکنده شده بودند. ماهیگیران محلی هر روز با فاچه‌های چوبی خود به میان آبها می‌رفتند و با تورهای دست‌بافشان، ماهی‌های فراوانی صید می‌کردند. این دریاچه نه تنها منبع غذایی، بلکه هویت و حیات اقتصادی جوامع اطراف بود.

اما با گذشت زمان، رودخانه‌هایی که قرن‌ها این دریاچه را سیراب می‌کردند، به سوی زمین‌های کشاورزی جدید منحرف شدند. کمک آب دریاچه کاهش یافت و خط ساحلی آن هر سال بیشتر از قبیل عقب‌نشینی کرد. ماهی‌ها ناپدید شدند و قلیق‌های ماهیگیری بر روی شن‌زارهای خشک رها شدند. مردم محلی که زمانی با غرور به دریاچه وابسته بودند، حالا با چهره‌های غمگین به زمین‌های خشک شده نگاه می‌کردند. بادهای شدید، نمک و مواد سمی را از بستر خشک دریاچه به وا بلند می‌کرد و بیماری‌های تفتیس میان کودکان و سالمدان شیوع پیدا کرد.

سال‌ها بعد، دانشمندان و مهندسان تلاش کردند تا بخشی از دریاچه را با ساخت یک سد احیا کنند. آب کمک بازگشت و ماهی‌ها دوباره در بخش‌هایی که کوچکی از دریاچه ظاهر شدند. اما بیشتر آنچه زمانی چهارمین دریاچه بزرگ جهان بود، اکنون به بیاناتی نمکی تبدیل شده بود که تنها خاطره‌ای از شکوه گذشته را در دل خود نگاه داشته بود. جهان این تراژی زیست‌محیطی را به عنوان یکی از بزرگترین فحایع طبیعی ثبت کرد، اما برای سیاری از مردم محلی، این فقط یک داستان غمگین از دست رفتن خانه‌ای بود که هرگز به طور کامل باز نگشت.

Rate

B

Chatgpt:

Multiple Choice Question: چرا مردم محلی به دریاچه «دریای جز ایر» می‌گفتهند؟ A) به خاطر عق زیاد آن B) به خاطر وجود هزاران جزیره در آن C) به خاطر آب شور آن D) به خاطر ماهی‌های فراوان آن Answer: B	A
True/False Judgment 1: مردم محلی پس از خشک شدن دریاچه، همچنان به ماهیگیری ادامه دادند Answer: نادرست (False)	A
True/False Judgment 2: خشک شدن دریاچه باعث افزایش بیماری‌های تنفسی در نظرفه شد. Answer: درست (True)	A
Cause and Effect: Cause: روندخانه‌هایی که دریاچه را سیراب می‌کرند به سوی زمین‌های کشاورزی منحرف شدند. Effect: آب دریاچه کاهش یافت و خط ساحلی آن عقب‌نشینی کرد.	A
Cloze: Sentence: بادهای شدید، نمک و مواد سمی را از بستر خشک دریاچه به هوا بلند می‌کرد و «_____» میان کوکان و سالمدان شیوع پیدا کرد Answer: بیماری‌های تنفسی	A
Explanation: Question: چرا مردم محلی پس از خشک شدن دریاچه غمگین بودند؟ Answer: زیرا دریاچه ناشتها منبع غذایی، بلکه هویت و حیات اقتصادی آن‌ها بود و از دست دادن آن به معنای از دست دادن خانه و میشتنشان بود	A
Counterfactual: Question: اگر روندانه‌ها به سوی زمین‌های کشاورزی منحرف نمی‌شدند، چه اتفاقی می‌افتد؟ Answer: دریاچه احتمالاً خشک نمی‌شد و مردم محلی همچنان می‌توانستند از منابع آن استفاده کنند	A

Chatgpt:

The Aral Sea (also known as *Kharazm Lake*) was a saltwater lake located in Central Asia, between Karakalpakstan in southern Uzbekistan and Kazakhstan to the north. Millions of years ago, the Aral Sea was part of the ancient Paratethys Sea. It was an endorheic (landlocked) salt lake that began shrinking in the 1960s and was largely dried up by the 2010s. The Aral Sea lay across Kazakhstan's Aktobe and Kyzylorda regions, and the autonomous Karakalpakstan region of Uzbekistan. Its name, which comes from Mongolic and Turkic roots, roughly translates to "Sea of Islands," a reference to the more than 1,100 islands that once dotted its waters. The Aral Sea's drainage basin covers parts of Uzbekistan, Afghanistan, Iran, Kazakhstan, Kyrgyzstan, Tajikistan, and Turkmenistan. Once the fourth-largest lake in the world, with an area of 68,000 square kilometers (26,300 square miles), the Aral Sea began to shrink in the 1960s when the rivers feeding it were diverted for Soviet irrigation projects. By 2007, the lake had shrunk to only 10% of its original size and had split into four separate bodies: the North Aral Sea, the eastern and western basins of the now much smaller South Aral Sea, and the small Barsakelmes Lake in the middle. By 2009, the southeastern lake had disappeared entirely, and the southwestern part had retreated into a narrow strip along the western edge of the former South Aral Sea. In the years that followed, occasional inflows of water have sporadically refilled parts of the southeastern lake to a small extent. NASA satellite images from August 2014 showed that, for the first time in modern history, the eastern basin of the Aral Sea had completely dried up. This area is now known as the Aralkum Desert. As part of Kazakhstan's efforts to save and restore the North Aral Sea, the Kokaral Dam was completed in 2005. By 2008, the water level had risen by 12 meters (39 feet) compared to 2003. By 2013, salinity had decreased, and fish populations had returned in sufficient numbers to allow some fishing activities to resume. After visiting the city of Muyнак in 2011, former UN Secretary-General Ban Ki-moon called the shrinking of the Aral Sea "one of the planet's worst environmental disasters." The once-thriving fishing industry has been destroyed, leading to unemployment and economic hardship in the region. Water from the Syr Darya River is now used to irrigate about 2 million hectares (5 million acres) of farmland in the Fergana Valley. The Aral Sea region is now heavily polluted, resulting in serious public health challenges. UNESCO has added historical records related to the Aral Sea to its Memory of the World Register as a source for studying this environmental tragedy.

Text	Rate
<p>Long ago, a vast saltwater lake once lay in the heart of Central Asia. The local people called it the "Sea of Islands" because of the thousands of small and large islands scattered across its shimmering surface like pearls. Every day, local fishermen would row their wooden boats into the waters, casting hand-woven nets and hauling in an abundance of fish. This lake was not only a source of food, but also the cultural identity and economic lifeblood of the surrounding communities. Over time, however, the rivers that had nourished the lake for centuries were diverted toward newly developed farmlands. Gradually, the lake's water levels declined, and its shoreline receded further each year. The fish disappeared, and fishing boats were left stranded on dry, sandy plains. The people, once proud and deeply connected to the lake, could only watch with sorrow as it turned to desert. Strong winds swept salt and toxic substances from the dry lakebed into the air, spreading respiratory illnesses among children and the elderly. Years later, scientists and engineers tried to revive part of the lake by constructing a dam. Slowly, water returned to some areas, and fish began to reappear in small sections. But most of what was once the fourth-largest lake in the world had become a salty desert — holding only the memory of its former glory. The world came to recognize this environmental tragedy as one of the greatest ecological disasters in modern history. Yet, for many of the local people, it remained a deeply personal loss — the story of a home that was never truly restored.</p>	B

Chatgpt:

Multiple Choice Question: Why did the locals call the lake the “Sea of Islands”? A) Because of its great depth. B) Because it contained thousands of islands. C) Because of its salty water. D) Because of its abundance of fish. Answer: B	A
True/False Judgment 1: The locals continued fishing even after the lake dried up. Answer: False	A
True/False Judgment 2: The drying of the lake led to an increase in respiratory illnesses in the region. Answer: True	A
Cause and Effect: Cause: The rivers that once fed the lake were diverted toward agricultural lands. Effect: <u>The lake's water levels dropped, and its shoreline receded.</u>	A
Cloze: Sentence: Strong winds lifted salt and toxic substances from the dry lakebed into the air, and "_____" spread among children and the elderly. Answer: respiratory diseases	A
Explanation: Question: Why were the locals saddened after the lake dried up? Answer: Because the lake was not only a source of food but also a part of their identity and economic livelihood, and losing it meant losing their home and way of life.	A
Counterfactual: Question: What would have happened if the rivers hadn't been diverted toward farmland? Answer: The lake likely would not have dried up, and local people could have continued using its resources.	A

Qwen:

<p>مدیر عامل اجرایی (به انگلیسی: Chief Executive Officer) معمولاً عالیرتبه‌ترین مسئول اداری (اجرایی) با رئیس عهددار مدیریت تام یک شرکت سهامی، سازمان، یا یک بنگاه اقتصادی است که به هیئت مدیره گزارش می‌دهد. بهطور معمول یک مدیر ارشد اجرایی، چند مدیر اجرایی تابع دارد، که هر یک مسئولیت‌های کاربردی و پردازش خود را دارند. اعضای انجمن شامل مدیر ارشد عامل (COO)، مدیر ارشد فنی (CTO)، مدیر ارشد مالی (CFO)، مدیر ارشد بازرگانی (CCO)، مدیر ارشد بازاریابی (CMO)، مدیر ارشد فروش (CSO)، مدیر ارشد اطلاعاتی (CIO)، مدیر ارشد ارتباطات یا مدیر روابط عمومی (CCO)، مدیر ارشد اداری (CAO)، مدیر ارشد منابع انسانی (CHRO)، مدیر ارشد تدارکات (CPO) و یک مدیر ارشد آموزش (CLO).</p>	
<p>در یک شرکت بزرگ فناوری، محمد به عنوان مدیر ارشد اجرایی (CEO) تعیین شده بود و مسئولیت کلی رهبری و تصمیم‌گیری در شرکت را بر عهده داشت. او هر روز با هیئت مدیره جلسه داشت تا وضعیت شرکت را بررسی و برنامه‌های آینده را طراحی کند. تحت اداره محمد، تیمی از مدیران ارشد تخصصی فعالیت می‌کردند که هر کدام پوچه خاصی از کار شرکت را مدیریت می‌کردند.</p> <p>علی، مدیر ارشد عامل (COO)، مسئول بهینه‌سازی عملیات داخلی شرکت بود. حسن، مدیر ارشد فنی (CTO)، روزی پرورش‌های نوآورانه فناوری تمرکز داشت و سعی می‌کرد آخرین فناوری‌های را در محصولات شرکت استفاده کند. مریم، مدیر ارشد مالی (CFO)، بودجه‌بندی و مدیریت مالی شرکت را بر عهده داشت و اطمینان می‌گرفت که هزینه‌ها تحت کنترل باشند.</p>	<p>Rate</p>
<p>همچنین، سیمیه به عنوان مدیر ارشد بازاریابی (CMO) نقش مهمی در تبلیغ محصولات شرکت ایفا می‌کرد و استراتژی‌های بازاریابی جذابی طراحی می‌کرد. احمد، مدیر ارشد فروش (CSO)، با مشتریان همکاری می‌کرد تا بیشترین سود را برای شرکت بدست آورد. از سوی دیگر، فاطمه به عنوان مدیر ارشد منابع انسانی (CHRO) به توسعه و آموزش نیروی کار می‌پرداخت و سعی داشت محیط کاری شرکت را بهتر کند.</p> <p>تماس مستقیم این مدیران با محمد، CEO شرکت، باعث می‌شد که تصمیم‌گیری‌ها به سرعت و با دقت انجام شوند. همه این تیم‌ها با همکاری چشمگیر، شرکت را به سمت موفقیت‌های بزرگ رسانده بودند.</p>	<p>B</p>

Qwen:

<p>Multiple Choice Question: چرا تماس مستقیم مدیران ارشد با محمد، CEO شرکت، در تصمیم‌گیری‌های سریع و دقیق نقش دارد؟ A) چون همه مدیران از تجربه کمتری برخوردارند B) چون اطلاعات به صورت فوری و بیون فیلتر به رهبر منتقل می‌شود C) چون محمد به عنوان تنها نفری که تصمیم‌گیری می‌کند عمل می‌کند D) چون مدیران ارشد باید فقط تحت فرمان او کار کنند</p> <p>Answer: B</p>	A
<p>True/False Judgment 1: تماس مستقیم مدیران ارشد با CEO، سرعت تصمیم‌گیری را افزایش می‌دهد. Answer: (True)</p>	A
<p>True/False Judgment 2: محمد، CEO، تنها شخصی است که می‌تواند تصمیمات مهم شرکت را بگیرد. Answer: (False)</p>	A
<p>Cause and Effect: Cause: تماس مستقیم مدیران ارشد با محمد، CEO شرکت Effect: تصمیم‌گیری‌های سریع و دقیق انجام می‌شود.</p>	A
<p>Cloze: Sentence: تحت اداره محمد، نیمی از مدیران ارشد تخصصی فعالیت می‌کردند که هر کدام «_____ خاصی از کار شرکت را مدیریت می‌کردند. Answer: حوزه‌ای</p>	B
<p>Explanation: Question: چرا مدیران ارشد مختلف در این شرکت نقش‌های متفاوتی ایفا می‌کنند؟ Answer: مدیران ارشد به منظور تخصص‌بندی و بهینه‌سازی عملکرد شرکت، هر کدام در حوزه‌ای مشخص فعالیت می‌کنند.</p>	A
<p>Counterfactual: Question: اگر تماس مستقیم مدیران ارشد با محمد قطع می‌شد، چه اتفاقی می‌افتد؟ Answer: تصمیم‌گیری‌ها طولانی‌تر می‌شد و احتمالاً کیفیت آنها کاهش می‌یافتد.</p>	A

Qwen:

The Chief Executive Officer (CEO) is typically the highest-ranking administrative (executive) authority within a corporation, organization, or business entity, and is responsible for overall management. The CEO reports directly to the board of directors. Usually, the CEO oversees several subordinate executive officers, each responsible for a specific functional area within the organization. These roles often include the Chief Operating Officer (COO), who manages day-to-day operations; the Chief Technology Officer (CTO), responsible for technological direction; and the Chief Financial Officer (CFO), who handles the financial strategy and planning. Other key positions may include the Chief Commercial Officer (CCO), Chief Marketing Officer (CMO), Chief Sales Officer (CSO), and Chief Information Officer (CIO), each managing their respective departments. Additionally, organizations often have a Chief Communications Officer or Public Relations Director, a Chief Administrative Officer (CAO), a Chief Human Resources Officer (CHRO), a Chief Procurement Officer (CPO), and a Chief Learning Officer (CLO), who oversees training and development. Each of these executives plays a critical role in the strategic and operational success of the organization, working under the leadership and direction of the CEO.

In a large technology company, Mohammad was appointed as the Chief Executive Officer (CEO), bearing the overall responsibility for leadership and decision-making within the organization. Every day, he held meetings with the board of directors to review the company's status and plan future strategies.	Rate
<p>Under Mohammad's leadership, a team of specialized senior executives managed various key areas of the company. Ali, the Chief Operating Officer (COO), was responsible for optimizing the company's internal operations. Hassan, the Chief Technology Officer (CTO), focused on innovative technology projects, aiming to integrate the latest technologies into the company's products. Maryam, the Chief Financial Officer (CFO), oversaw budgeting and financial management, ensuring that expenses remained under control. Somayeh, serving as the Chief Marketing Officer (CMO), played a vital role in promoting the company's products and designing engaging marketing strategies. Ahmad, the Chief Sales Officer (CSO), worked closely with clients to maximize company profits. Meanwhile, Fatemeh, as the Chief Human Resources Officer (CHRO), focused on workforce development and training, striving to improve the workplace environment. The direct communication between these executives and Mohammad, the CEO, enabled fast and accurate decision-making. Through this remarkable collaboration, the entire team had led the company toward major successes.</p>	B

Qwen:

<p>Multiple Choice Question: Why does direct contact between senior managers and Mohammad, the company's CEO, contribute to fast and accurate decision-making? A) Because all managers have less experience. B) Because information is delivered to the leader instantly and without filtering. C) Because Mohammad acts as the sole decision-maker. D) Because senior managers must only work under his command.</p> <p>Answer: B</p>	A
<p>True/False Judgment 1: Direct contact between senior managers and the CEO increases the speed of decision-making.</p> <p>Answer: True</p>	A
<p>True/False Judgment 2: Mohammad, the CEO, is the only person who can make important decisions in the company.</p> <p>Answer: False</p>	A
<p>Cause and Effect: Cause: Direct contact between senior managers and Mohammad, the CEO of the company. Effect: Decisions are made quickly and accurately.</p>	A
<p>Cloze: Sentence: Below eye of Mohammad's leadership, a team of specialized senior managers worked, each managing a specific "____" of the company's operations.</p> <p>Answer: area</p>	B
<p>Explanation: Question: Why do different senior managers in this company play different roles?</p> <p>Answer: Senior managers work in specific areas to specialize and optimize the company's performance.</p>	A
<p>Counterfactual: Question: What would happen if direct contact between senior managers and Mohammad was cut off?</p> <p>Answer: Decision-making would take longer and probably decrease in quality.</p>	A

Mistral:

<p>بیمارستان، بهداری، درمانگاه، شفاخانه یا مریضخانه (به انگلیسی: Hospital) یک بنیاد مرافقی های بهداشتی است که با کارکنان پزشکی و پرستاری و تجهیزات پزشکی به بیماران خدمات درمانی ارائه می دهد. پرآواز هنرین گونه بیمارستان، بیمارستان عمومی است که بهطور معمول دارای یک بخش اورژانس برای درمان مشکلات فوری سلامتی از قربانیان آتشسوزی و تصادف گرفته تا بیماری های ناگهانی است.</p> <p>یک بیمارستان منطقه‌ای بهطور معمول مهمترین مرکز مرافقی های بهداشتی در منطقه خود است، که دارای تعداد زیادی تخت برای مرافقی های ویژه و تخت های اضافی برای بیمارانی است که نیاز به مرافقی طولانی مدت دارند.</p> <p>بیمارستان های تخصصی شامل مراکز ترموما، بیمارستان های توانبخشی، بیمارستان بزرگسالان (سالمندان) و بیمارستان هایی برای رسیدگی به نیاز های خاص پزشکی مانند درمان بیماری های روانی (مراجعه به بیمارستان روانپزشکی) و برخی از دسته های بیماری هستند. بیمارستان های تخصصی می توانند در مقایسه با بیمارستان های عمومی به کاهش هزینه های مرافقی های بهداشتی کمک کنند. بیمارستان های بسته به منبع درآمد به عمومی، تخصصی یا دولتی رده بندی می شوند.</p> <p>یک بیمارستان آموزشی کمک به افراد را با آموزش به دانشجویان پزشکی و پرستاران ترکیب می کند. عموماً به یک مرکز درمانی کوچکتر از بیمارستان، درمانگاه یا کلینیک گفته می شود. بیمارستان ها دارای طیف گسترده ای از بخش ها (به عنوان مثال جراحی و مرافقی های فوری) و واحدهای تخصصی مانند قلب و عروق هستند. بعضی از بیمارستان ها بخش های سرپایی دارند و برخی دیگر واحدهای درمانی مزمن دارند. واحدهایی که از بیمارستان ها پشتیبانی می کنند معمولاً شامل داروخانه ها، آزمایشگاه های پاتولوژی و رادیولوژی هستند.</p> <p>بودجه بیمارستان ها بهطور معمول از راه تأمین بودجه عمومی، سازمان های بهداشتی (با هدف انتفاعی یا غیر انتفاعی)، شرکت های بیمه درمانی یا نهادهای نیکوکاری از جمله کمک های مستقیم تأمین می شود. در گذشته، بیمارستان ها بیشتر با دستور ها مذهبی، یا افراد و رهبران نیکوکار تأسیس و تأمین می شدند.</p>	<p>علی که دانشجوی پزشکی بود، هر روز صبح زود به بیمارستان آموزشی می رفت تا در کنار پزشکان مجرب و پرستاران، تجربیات جدیدی کسب کند. این بیمارستان که بزرگترین مرکز مرافقی های بهداشتی در منطقه بود، دارای تعداد زیادی تخت برای مرافقی های ویژه و تخت های اضافی برای بیمارانی که نیاز به مرافقی طولانی مدت داشتند. علی هر روز در بخش های مختلف مانند جراحی، مرافقی های فوری و واحدهای تخصصی مانند قلب و عروق حضور داشت.</p> <p>یک روز، در حالی که علی در بخش اورژانس حضور داشت، یک مورد فوری به بیمارستان آورده شد. قربانی یک تصادف جاده ای بود که به شدت مجرح شده بود. تیم پزشکی بالاگصلة به کمک او شفافتد و علی نیز به عنوان یک دانشجوی پزشکی، به آنها کمک می کرد. این تجربه برای علی بسیار آموزنده بود، زیرا او برای اولین بار با مواجهه با یک مورد واقعی فوری، تجربه ای مستقیم از کار در بخش اورژانس به دست آورد.</p> <p>بیمارستانی که علی در آن تحصیل می کرد، یک بیمارستان عمومی بود که توسط بودجه های عمومی و کمک های مستقیم از سازمان های بهداشتی و نهادهای نیکوکاری تأمین می شد. این بیمارستان هچنین دارای واحدهای پشتیبانی مانند داروخانه ها، آزمایشگاه های پاتولوژی و رادیولوژی بود که به پزشکان و بیماران کمک می کردند.</p> <p>علی هر روز بالخندی روی چهره به بیمارستان می رفت، زیرا می داشت که هر روز یادگیری های جدیدی در انتظار او است. او بالا نشود و پشتکار خود روز به روز به یک پزشک ماهر تبدیل می شد و آماده می شد تا در اینده به جامعه خدمت کند.</p>
<p>علی که دانشجوی پزشکی بود، هر روز صبح زود به بیمارستان آموزشی می رفت تا در کنار پزشکان مجرب و پرستاران، تجربیات جدیدی کسب کند. این بیمارستان که بزرگترین مرکز مرافقی های بهداشتی در منطقه بود، دارای تعداد زیادی تخت برای مرافقی های ویژه و تخت های اضافی برای بیمارانی که نیاز به مرافقی طولانی مدت داشتند. علی هر روز در بخش های مختلف مانند جراحی، مرافقی های فوری و واحدهای تخصصی مانند قلب و عروق حضور داشت.</p> <p>یک روز، در حالی که علی در بخش اورژانس حضور داشت، یک مورد فوری به بیمارستان آورده شد. قربانی یک تصادف جاده ای بود که به شدت مجرح شده بود. تیم پزشکی بالاگصلة به کمک او شفافتد و علی نیز به عنوان یک دانشجوی پزشکی، به آنها کمک می کرد. این تجربه برای علی بسیار آموزنده بود، زیرا او برای اولین بار با مواجهه با یک مورد واقعی فوری، تجربه ای مستقیم از کار در بخش اورژانس به دست آورد.</p> <p>بیمارستانی که علی در آن تحصیل می کرد، یک بیمارستان عمومی بود که توسط بودجه های عمومی و کمک های مستقیم از سازمان های بهداشتی و نهادهای نیکوکاری تأمین می شد. این بیمارستان هچنین دارای واحدهای پشتیبانی مانند داروخانه ها، آزمایشگاه های پاتولوژی و رادیولوژی بود که به پزشکان و بیماران کمک می کردند.</p> <p>علی هر روز بالخندی روی چهره به بیمارستان می رفت، زیرا می داشت که هر روز یادگیری های جدیدی در انتظار او است. او بالا نشود و پشتکار خود روز به روز به یک پزشک ماهر تبدیل می شد و آماده می شد تا در اینده به جامعه خدمت کند.</p>	<p>Rate</p> <p>B</p>

Mistral:

Multiple Choice Question: علی به چه دلیل هر روز بالبختی روی چهره به بیمارستان میرفت؟ A) به دلیل دریافت حقوق بالا B) به دلیل یادگیری های جدید و تجربیات ارزشمند C) به دلیل دوست داشتن غذای بیمارستان D) به دلیل نزدیکی به خانه Answer: B	C
True/False Judgment 1: علی به دلیل دریافت حقوق بالا هر روز بالبختی روی چهره به بیمارستان میرفت. Answer: (False)	A
True/False Judgment 2: علی به دلیل یادگیری های جدید و تجربیات ارزشمند هر روز بالبختی روی چهره به بیمارستان میرفت. Answer: (True)	C
Cause and Effect: علی هر روز در بخش های مختلف بیمارستان حضور داشت. Cause: علی تجربیات جدیدی کسب کرد و به یک پزشک ماهر تبدیل شد. Effect: علی هر روز با «_____» روی چهره به بیمارستان میرفت.	A
Cloze: علی هر روز با «_____» روی چهره به بیمارستان میرفت. Sentence: علی هر روز بالبختی روی چهره به بیمارستان میرفت. Answer: لبخندی	C
Explanation: Question: چرا علی هر روز بالبختی روی چهره به بیمارستان میرفت؟ Answer: علی به دلیل علاقه اش به یادگیری و کسب تجربیات جدید هر روز بالبختی روی چهره به بیمارستان میرفت.	C
Counterfactual: Question: اگر علی هر روز بالبختی روی چهره به بیمارستان نمیرفت، چه می شد؟ Answer: علی احتسالاً انگیزه اش برای یادگیری و کسب تجربیات جدید کمتر می شد و احتسالاً به یک پزشک ماهر تبدیل نمی شد.	C

Mistral:

A hospital (also known as a medical center, clinic, infirmary, or health facility) is a healthcare institution that provides treatment to patients through medical staff, nurses, and medical equipment. The most well-known type is the general hospital, which typically includes an emergency department for treating urgent health issues—from burn and accident victims to sudden illnesses.

A regional hospital is usually the main healthcare facility in its area, equipped with numerous intensive care beds and additional beds for patients requiring long-term care.

Specialized hospitals include trauma centers, rehabilitation hospitals, children's hospitals, geriatric (elderly) hospitals, and institutions for specific medical needs—such as psychiatric hospitals for mental health conditions or facilities dedicated to certain disease categories. Compared to general hospitals, specialized hospitals can help reduce healthcare costs.

Hospitals are categorized as public, private, or governmental depending on their funding sources.

A teaching hospital combines patient care with the education of medical students and nurses.

Smaller medical facilities than hospitals are usually referred to as clinics or health centers.

Hospitals typically include a wide range of departments (such as surgery and emergency care) and specialized units like cardiology.

Some hospitals also have outpatient departments, while others maintain chronic care units.

Supporting units usually include pharmacies, pathology labs, and radiology departments.

Hospital funding commonly comes from public funding, health organizations (whether for-profit or non-profit), health insurance companies, or charitable donations, including direct contributions.

Historically, many hospitals were founded and supported by religious orders, philanthropists, or benevolent leaders.

Ali, a medical student, went to the teaching hospital early every morning to gain new experiences alongside experienced doctors and nurses. This hospital, the largest healthcare center in the region, had a large number of beds for intensive care as well as additional beds for patients requiring long-term treatment. Each day, Ali rotated through various departments such as surgery, emergency care, and specialized units like cardiology.

One day, while Ali was stationed in the emergency department, a critical case was brought in. The patient was a road accident victim who was severely injured. The medical team immediately rushed to assist, and Ali, as a medical student, joined in to help. The experience was deeply educational for Ali, as it was his first time directly involved in managing a real emergency case.

The hospital where Ali was training was a public hospital, funded through public budgets and direct support from health organizations and charitable institutions. It also had supporting units such as pharmacies, pathology labs, and radiology departments that assisted both doctors and patients.

Every day, Ali arrived at the hospital with a smile on his face, knowing that new lessons and insights awaited him. Through dedication and persistence, he was steadily becoming a skilled physician, preparing himself to serve society in the future.

Rate

B

Mistral:

Multiple Choice Question: Why did Ali go to the hospital every day with a smile on his face? A) Because he was getting a high salary B) Because of the new learning opportunities and valuable experiences C) Because he liked the hospital food D) Because it was close to his home Answer: B	C
:True/False Judgment 1 Ali went to the hospital every day with a smile on his face because he was receiving a high salary :Answer False (False)	A
True/False Judgment 2: Ali went to the hospital every day with a smile on his face because of the new learning opportunities and valuable experiences. Answer: True(True)	C
Cause and Effect: Cause: Ali was present in various departments of the hospital every day. Effect: Ali gained new experiences and became a skilled doctor.	A
Cloze: Sentence: Every day, Ali went to the hospital with a « ____ » on his face. Answer: smile	C
Explanation Question: Why did Ali go to the hospital every day with a smile on his face? Answer: Because he was passionate about learning and gaining new experiences, Ali went to the hospital every day with a smile on his face.	C
Counterfactual Question: What would have happened if Ali didn't go to the hospital every day with a smile on his face? Answer: Ali would probably have been less motivated to learn and gain new experiences, and he might not have become a skilled doctor.	C

Gemini:

Rate	A
۱۲	۳/۵

Gemini:

<p>Multiple Choice Question:</p> <p>چرا بزرگان محله بر اهمیت راهکارهای پیشگیری پس از وقوع این فاجعه تأکید کرند؟ زیرا معنف بودند که این فاجعه اختیابناپذیر بود</p> <p>A) برای اطمینان از اینکه احتمال وقوع حوادث مشابه در آینده کاهش باید. B) تا فردی که خودکشی کرده بود را سرزنش کنند C) برای جماعت این امر جهت گزارش‌های جهانی D) برای جماعت این امر جهت گزارش‌های جهانی</p> <p>Answer: B</p>	A
<p>True/False Judgment 1:</p> <p>فقط یک علت واحد و مشخص برای تمايل افراد به پایان دادن به زندگی خود وجود ندارد.</p> <p>Answer: درست (True)</p>	A
<p>True/False Judgment 2:</p> <p>میزان مرگ ناشی از خودکشی در زنان به مراتب بیشتر از مردان است.</p> <p>Answer: نادرست (False)</p>	A
<p>Cause and Effect:</p> <p>Cause: افزایش آگاهی عمومی و درمان به موقع اختلالات روانی</p> <p>Effect: کاهش چشمگیر آمار خودکشی در جامعه</p>	A
<p>Cloze:</p> <p>Sentence: در آمریکا، خطوط تلفن بحران رایج است، گرچه اثربخشی آن‌ها کاملاً «_____» است</p> <p>Answer: اثبات نشده</p>	A
<p>Explanation:</p> <p>Question: چرا محدود کردن دسترسی به ابزارهای خطرناک یکی از راهکارهای مهم برای پیشگیری از خودکشی محسوب می‌شود؟</p> <p>Answer: با کاهش یا از بین بردن ابزارهای مرگبار، احتمال موفقیت در اقامه به خودکشی به طور قابل توجهی کاهش می‌یابد و فرد فرصت بیشتری برای پشیمانی یا دریافت کمک پیدا می‌کند</p>	A
<p>Counterfactual:</p> <p>Question: چه اتفاقی می‌افتد اگر اهالی شهر پس از شنیدن خبر فاجعه، به جای همبستگی و آگاهی‌بخشی، این موضوع را نادیده می‌گرفتند؟</p> <p>Answer: اگر جامعه این موضوع را نادیده می‌گرفت، احتمالاً آگاهی عمومی درباره علل و راه‌های پیشگیری از خودکشی افزایش نمی‌یافتد و خطر تکرار چنین فاجعی در آینده همچنان بالا می‌ماند.</p>	A

Gemini:

range of contributing factors. Feelings of hopelessness, mental disorders such as depression, bipolar disorder, schizophrenia, and autism spectrum disorders, as well as physical conditions like chronic fatigue syndrome, romantic breakups, alcoholism, and substance or drug abuse, have all been linked to increased suicide risk. Proven evidence shows that these factors can make individuals more vulnerable to suicide.

Stressful life events—such as financial problems, academic or job-related difficulties, interpersonal issues like divorce or breakups, abuse, or bullying—can also play a significant role.

People who have attempted suicide before are at a much higher risk of future attempts. Effective suicide prevention strategies include restricting access to lethal means (e.g., firearms, medications, and poisons), responsible and accurate media reporting on suicide, treatment of mental health and substance use disorders, and improving economic conditions. Although crisis hotlines are a common preventive resource in the United States, their effectiveness has not been thoroughly studied.

The most common method of suicide varies from country to country and often depends on the availability of lethal means. Typical methods include hanging, pesticide poisoning, and firearms. In 2015, suicide was responsible for 828,000 deaths worldwide—up from 712,000 in 1990—making it the tenth leading cause of death globally.

Approximately 1.5% of all global deaths are due to suicide. In a given year, this rate is around 12 per 100,000 people. Suicide rates are generally higher among men than women, ranging from 1.5 times higher in developing countries to 3.5 times higher in developed nations. Suicide is most common among people over the age of 70, although in some countries, those aged 15 to 30 are at the highest risk. As of 2015, Europe had the highest suicide rate by region. Each year, there are an estimated 10 to 20 million unsuccessful suicide attempts, which can often lead to long-term injury or disability. In the Western world, ...

The city had fallen into a heavy silence. A heartbreaking piece of news had gripped everyone's heart: a young man had taken his own life. This deliberate act leading to death was a painful reminder of a deeply rooted and complex phenomenon. The townspeople were well aware that despair, mental health disorders such as depression, physical illnesses, emotional heartbreak, and even addiction and substance abuse could place a person at serious risk—especially if they had previously attempted something similar. Everyday pressures only made things worse: financial struggles, work-related stress, family conflicts, or experiences of bullying and harassment.

One of the community elders reminded everyone, "We must start thinking about prevention. Limiting access to dangerous means—like hanging or agricultural poisons, which are more common here, or firearms in other regions—and also promoting responsible media coverage, are all vital steps. Of course, treating mental illnesses and improving economic conditions also play major roles." He added, "In the U.S., crisis hotlines are common, though their effectiveness has not been fully studied."

Global statistics were equally shocking: 828,000 suicide deaths in 2015, a dramatic increase from 712,000 in 1990. This had made suicide the 10th leading cause of death worldwide, accounting for about 1.5% of all global deaths, or approximately 12 people per 100,000 per year. It was also noted that suicide rates are generally higher among men than women, and while it is more prevalent in people over the age of 70, in some countries, young people aged 15 to 30 are at the highest risk. Europe had the highest regional suicide rate in 2015.

Rate

A

Gemini:

<p>Multiple Choice Question: Why did the community elders emphasize the importance of preventive measures after the tragedy occurred?</p> <p>A) Because they believed the tragedy was inevitable. B) To ensure the likelihood of similar incidents in the future would be reduced. C) To blame the person who had committed suicide. D) To collect statistics for global reports.</p> <p>Answer: B</p>	A
<p>True/False Judgment 1: There is not just a single, clear cause behind people's desire to end their lives.</p> <p>Answer: True</p>	A
<p>True/False Judgment 2: The suicide death rate is significantly higher among women than men.</p> <p>Answer: False</p>	A
<p>Cause and Effect Cause: Increased public awareness and timely treatment of mental disorders Effect: Significant reduction in suicide rates in the community</p>	A
<p>Cloze: Sentence: In the United States, crisis hotlines are common, although their effectiveness is still " ____ ". Answer: unproven</p>	A
<p>Explanation Question: Why is limiting access to dangerous means considered an important strategy for suicide prevention? Answer: Because by reducing or eliminating access to deadly means, the likelihood of a successful suicide attempt decreases significantly, giving the person more time to reconsider or seek help.</p>	A
<p>Counterfactual Question: What would have happened if the townspeople had ignored the tragedy instead of coming together and raising awareness? Answer: If the community had ignored the incident, public awareness of the causes and prevention of suicide likely would not have increased, and the risk of similar tragedies recurring in the future would have remained high.</p>	A

Gemini:

نوروز نخستین روز سال خورشیدی برابر با یکم فروردین، جشن آغاز سال نوی ایرانی و یکی از کهن‌ترین جشن‌های به جا مانده از دوران ایران باستان است. خاستگاه نوروز در ایران باستان است و در مناطق وسیعی در آسیا و دیگر نقاط جهان جشن گرفته می‌شود. زمان برگزاری نوروز، اعتدال بهاری و در آغاز فصل بهار است. نوروز، با عنوان رسمی «روز بین‌المللی نوروز»، توسط یونسکو به عنوان میراث فرهنگی و معنوی بشر به ثبت جهانی رسیده است.

نوروز در ایران و افغانستان آغاز سال نو محسوب می‌شود و در برخی دیگر از کشورها یعنی تاجیکستان، روسیه، قرقیزستان، قرقیزستان، سوریه، عراق، گرجستان، جمهوری آذربایجان، آلبانی، چین، ترکمنستان، هند، پاکستان و ازبکستان تعطیل رسمی است و مردمان آن جشن را بربرا می‌کنند.

نوروز به عنوان یک جشن قبیلی و کهن از اقوام متعددی در منطقه به ارث رسیده است و بارنگ و انگ و پیزه، شناسنامه ایرانی به خود گرفته و از آن به عنوان نماد پیوند دهنده افراد و اقوام ایرانی و ملت ایران یاد می‌شود. ایرانیان نوروز را آغازگر رستاخیز طبیعت، گاه رویش و زایش باغ و بوستان می‌دانند و بر این باور هستند که در نوروز، همزمان با طبیعت، باید روزگار نو و جدیدی را با روان و نگرش نو، در تزییش تازه‌ای آغاز کنند. این روز همچنین یکی از روزهای مقدس و اعیاد مذهبی زرتشتیان و بهائیان به شمار می‌رود.

همین که هوای او اخیر اسفند ملایمتر شد و بوی گل‌های یاس به مشام رسید، نگاهی به تعقیم انداختم؛ اول فروردین، درست همزمان با اعتدال بهاری، رویش واژه‌ی «نوروز» را کنارش دیدم که نشان از آغاز سال نو خورشیدی می‌داد. برای ما ایرانیان، نوروز نه فقط یک جشن، بلکه جشن آغاز سال نوی ماست و پدربرگم همیشه از آن به عنوان یکی از کهن‌ترین جشن‌هایی یاد می‌کرد که از دل ایران باستان به ما به ارث رسیده است، ریشه‌هایش عیقاً در همین خاک نهفته. این جشن تنها مختص ماست؛ دوستم که اهل افغانستان است، هر سال همزمان با نوروز سال نو را تبریک می‌کوید. عجیب‌تر اینکه وقتی از دوستانم در اقصی نقاط جهان خبر گرفتم، دیدم که اقوام سیاری در مناطق وسیعی در آسیا و دیگر نقاط جهان، حتی در کشورهایی چون تاجیکستان، قرقیزستان و جمهوری آذربایجان آن را تعطیل رسمی می‌دانند و جشن می‌گیرند. شنیده بودم یونسکو نیز با عنوان رسمی «روز بین‌المللی نوروز»، آن را به عنوان میراث فرهنگی و معنوی بشریت به ثبت جهانی رسانده است. با این حال، نوروز بارنگ و بوی خاص خودش، هویت ایرانی به خود گرفته و نماد پیوند دهنده افراد و اقوام گوئانگون ایرانی است. ما ایرانی‌ها عیقاً باور داریم که نوروز آغازگر رستاخیز طبیعت، زمان رویش و زایش باغ و بوستان است. و درست مثل طبیعت، ما نیز باید با روانی تازه و نگاهی نو، با پوششی نو، روزگاری جدید را آغاز کنیم. تازه به یاد آوردم که نوروز برای دوستان زرتشتی و بهائی‌ان نیز یکی از مهم‌ترین روزهای مقدس و اعیاد مذهبی به شمار می‌رود. این شور و شوق مشترک، واقعاً هر ساله دیدنی است.

Rate

A

Gemini:

Multiple Choice Question: چرا ایرانیان در آستانه نوروز تمايل به خريد لباس ندارند؟ A) برای پیروی از یک مدد B) برای نشان دادن تجدید و شروعی تازه در سال نو C) به دلیل تغییرات آب و هوایی D) چون معمولاً لباس‌هایی کهنه از بین رفته‌اند Answer: B	A
True/False Judgment 1: جشن نوروز هواره در فصلی آغاز می‌شود که طبیعت شروع به بیدار شدن و رویش مجدد می‌کند. Answer: (True)	A
True/False Judgment 2: نوروز، به دلیل ریشه‌های عمیق در ایران باستان، تنها برای مردم ایران اهمیت مذهبی و فرهنگی دارد. Answer: (False)	A
Cause and Effect: Cause: یونسکو نوروز را به عنوان «روز بین‌المللی نوروز» و میراث فرهنگی و معنوی بشریت به ثبت جهانی رساند. Effect: این اقدام منجر به افزایش تأکید و تلاش برای حفظ و معرفی نوروز در سطح بین‌المللی شد.	A
Cloze: Sentence: ایرانی‌ها عیقاً باور دارند که همانطور که نوروز آغازگر رستاخیز طبیعت است، برای انسان‌ها نیز زمان مناسبی برای «_____» است. Answer: تجدد روحی و شروعی تازه	A
Explanation: Question: چرا نویسنده از عبارت «شور و شوق مشترک» برای توصیف حس نوروز استفاده می‌کند؟ Answer: زیرا متن نشان می‌دهد که نوروز نه تنها توسط ایرانیان با مذهب و اقوام مختلف جشن گرفته می‌شود، بلکه کشور‌های دیگری نیز در این جشن سهیم هستند و این اشتباه باعث ایجاد حس همیستگی و شادی جمعی می‌شود.	A
Counterfactual: Question: اگر نوروز با اعتدال پاییزی مصادف می‌شد و آغازگر فصل سرما بود، چه تاثیری بر نادگرایی آن می‌گذاشت؟ Answer: در این صورت، نوروز به جای نماد «رویش و زایش طبیعت» و «تجدد حیات»، بیشتر می‌توانست نماد «خذان» و «آماده شدن برای استراحت» باشد که معنای فعلی آن را تغییر می‌داد.	A

Gemini:

Nowruz is the first day of the solar year, corresponding to the first day of *Farvardin* (the first month in the Iranian calendar). It marks the celebration of the Iranian New Year and is one of the oldest surviving festivals from ancient Persia. The origin of Nowruz lies in ancient Iran, and it is celebrated across vast regions of Asia and other parts of the world. Nowruz is held during the vernal equinox and marks the beginning of spring. Nowruz is considered the start of the new year in Iran and Afghanistan, and it is also a public holiday in several other countries including Tajikistan, Russia, Kyrgyzstan, Kazakhstan, Syria, Iraq, Georgia, Azerbaijan, Albania, China, Turkmenistan, India, Pakistan, and Uzbekistan, where people celebrate it with joy. As an ancient and traditional celebration inherited from many ethnic groups in the region, Nowruz has acquired a distinct Iranian identity and is regarded as a symbol that unites the Iranian people and ethnicities. Iranians see Nowruz as the beginning of nature's resurrection—a time of growth and rebirth in gardens and orchards—and believe that, just like nature, one should begin a new chapter in life with a renewed spirit, fresh mindset, and new clothes. This day is also considered a sacred occasion and a religious holiday among Zoroastrians and members of the Bahá'í faith.

As the late Esfand (the last month of the Persian calendar) air grew milder and the scent of jasmine flowers began to fill the air, I glanced at the calendar. The first of Farvardin had arrived—right at the spring equinox—and next to it was the word “Nowruz,” marking the beginning of the new solar year. For us Iranians, Nowruz is not just a celebration, but the true start of our new year. My grandfather always spoke of it as one of the most ancient festivals passed down from the depths of ancient Persia, with roots deeply embedded in this land. This celebration isn't only ours. My friend from Afghanistan greets me every year at Nowruz to mark the start of their new year as well. What surprised me even more was learning from friends around the world that many communities across vast regions of Asia and other parts of the world—including countries like Tajikistan, Kazakhstan, and Azerbaijan—celebrate it as an official holiday. I had heard that UNESCO officially recognized Nowruz as an “International Day of Nowruz,” registering it as part of the world's intangible cultural heritage. Still, Nowruz—with its unique colors, scents, and spirit—has taken on a distinctly Iranian identity, standing as a symbol of unity among diverse Iranian communities and ethnic groups. We Iranians deeply believe that Nowruz signals the rebirth of nature, the time when gardens and orchards bloom once again. Just like nature, we too are meant to begin a fresh chapter—with renewed spirit, a fresh perspective, and new clothing. I suddenly remembered that Nowruz is also one of the most sacred religious holidays for my Zoroastrian and Bahá'í friends. This shared excitement and energy each year is truly something to behold.

Rate

A

Gemini:

Multiple Choice Question: Why do Iranians tend to buy new clothes in the days leading up to Nowruz? A) To follow a new fashion trend. B) To symbolize renewal and a fresh start in the new year. C) Because of weather changes. D) Because old clothes are usually worn out. Answer: B	A
True/False Judgment 1: Nowruz always begins in a season when nature begins to awaken and grow again. Answer: True	A
True/False Judgment 2: Due to its deep roots in ancient Iran, Nowruz holds religious and cultural significance only for Iranians. Answer: False	A
Cause and Effect Cause: UNESCO registered Nowruz as the "International Day of Nowruz" and recognized it as an intangible cultural heritage of humanity. Effect: This led to increased global awareness and efforts to preserve and promote Nowruz internationally.	A
Cloze Sentence: Iranians deeply believe that just as Nowruz marks the rebirth of nature, it is also the right time for humans to experience spiritual renewal and a fresh beginning.	A
Explanation Question: Why does the author use the phrase "shared excitement" to describe the feeling around Nowruz? Answer: Because the text shows that Nowruz is celebrated not only by Iranians of various religions and ethnicities, but also in other countries, creating a sense of unity and collective joy.	A
Counterfactual Question: What would have happened if Nowruz had coincided with the autumnal equinox and marked the start of a colder season? Answer: In that case, Nowruz would likely be associated not with "rebirth and renewal of nature," but more with "decline" and "preparation for rest," thus changing its current symbolic meaning.	A

Mistral:

کوه سینا (به عربی: طور سیناء یا جبل موسی)، (به انگلیسی: Mount Sinai) (به عربی: جبل موسی به معنی خاردار) نام کوهی در شمال شرقی مصر است. این کوه در دین‌های ابراهمی بهویژه یهودیت شهرت فراوانی دارد. در روایت‌های سنتی این دین‌ها به این کوه اشاره بسیار شده‌است: کوه سینا که ایست که بنی اسرائیل سه ماه از دوران آوارگی ۴۰ ساله خود را آنجا سپری کردن بنی اسرائیل پس از ترک مصر به کوه سینا رسیدند و در آنجا اردو زدند. مطابق روایات سنتی ادیان ابراهمی خداوند شریعت را آنجا به بنی اسرائیل عطا کرد. گوسله طلایی سامری در این کوه ساخته شد. در این کوه یهود نخستین بار بر موسی وحی کرد و در بوته سوزان به صورت آتش بر وی تجلی نمود. و از خداوند به وی وحی شد که در سرزمین مقدس نعلین از پا بکند. این کوه حوریب، جبل الله، الطور، و جبل موسی نیز نامیده شده‌است. در قرآن به این کوه سوگند یاد شده‌است.

Rate	A
در یک روز گرم تابستانی، یک گروه کوچک از گردشگران ایرانی به شال شرقی مصر سفر کردند تا کوه سینا را ببینند. این کوه که در دین‌های ابراهمی بهویژه یهودیت شهرت فراوانی دارد، جایی است که بنی اسرائیل سه ماه از دوران آوارگی ۴۰ ساله خود را در آن سپری کردند. گردشگران با علاوه شنیدند که بنی اسرائیل پس از ترک مصر به این کوه رسیدند و در آنجا اردو زدند. راهنمای گردشگری به آنها توضیح داد که مطابق روایات سنتی ادیان ابراهمی، خداوند شریعت را در اینجا به بنی اسرائیل عطا کرده است. یکی از گردشگران با تعجب گفت: "ایا اینجا همان جایی است که گوسله طلایی سامری ساخته شده است؟" راهنمای با لبخندی پاسخ داد: "بله، همین است. اینجا همچنین جایی است که یهود نخستین بار بر موسی وحی کرد و در بوته سوزان به صورت آتش بر وی تجلی نمود." گردشگران با تعجب شنیدند که از خداوند به موسی وحی شده بود که در سرزمین مقدس نعلین از پا بکند. راهنمای اضافه کرد: "این کوه به نام‌های دیگری مانند حوریب، جبل الله، الطور، و جبل موسی نیز شناخته می‌شود. در قرآن نیز به این کوه سوگند یاد شده است." گردشگران با احساس تعجب و احترام، کوه سینا را ترک کردند، اما تجربه این سفر همیشه در ذهن‌شان باقی ماند.	

Mistral:

Multiple Choice Question: گردشگران در کوه سینا چه احساسی داشتند که این تجربه را همیشه در ذهنشان نگه داشتند؟ A) ترس B) تعجب و احترام C) خستگی D) نامیدی Answer: B	C
True/False Judgment 1: گردشگران به خاطر تعجب از تاریخ کوه سینا، این تجربه را همیشه در ذهنشان نگه داشتند. Answer: (True)	C
True/False Judgment 4: گردشگران به خاطر خستگی فیزیکی این تجربه را همیشه در ذهنشان نگه داشتند. Answer: (False)	A
Cause and Effect: Cause: گردشگران در مورد تاریخ و مقدس بودن کوه سینا آگاه شدند. Effect: گردشگران احساس تعجب و احترام کردند. Cloze: Sentence: گردشگران با احساس «_____» و احترام، کوه سینا را ترک کردند. Answer: تعجب	A
Explanation: Question: چرا گردشگران این تجربه را همیشه در ذهنشان نگه داشتند؟ Answer: به خاطر تعجب و احترامی که از تاریخ و مقدس بودن کوه سینا داشتند.	C
Counterfactual: Question: اگر گردشگران در مورد تاریخ کوه سینا آگاهی نداشتند، آیا آن را همیشه در ذهنشان نگه می‌داشتند؟ Answer: نه، چون تعجب و احترام از تاریخ و مقدس بودن مکان باعث می‌شود که افراد آن را همیشه در ذهنشان نگه دارند.	C

Mistral:

Mount Sinai (Arabic: جبل موسى or طور سيناء; English: Mount Sinai; Hebrew: הר סיני, meaning "thorny mountain") is the name of a mountain located in northeastern Egypt. It holds great significance in the Abrahamic religions, especially in Judaism. In traditional narratives of these religions, the mountain is frequently mentioned.

According to these traditions, Mount Sinai is the place where the Israelites spent three months during their 40-year period of wandering after leaving Egypt. Upon arriving at Mount Sinai, they set up camp there. It is traditionally believed that God gave the Law (the Torah) to the Israelites at this mountain. The Golden Calf of Samiri was also made at this site.

Mount Sinai is also known as the place where Yahweh first revealed Himself to Moses and appeared to him in the form of a burning bush. During this divine encounter, God commanded Moses to remove his sandals, as he was standing on holy ground. The mountain is also referred to as Horeb, Jabal Allah, al-Tur, and Jabal Musa.

In the Qur'an, God swears by this mountain, further emphasizing its sacredness in Islamic tradition.

On a hot summer day, a small group of Iranian tourists traveled to northeastern Egypt to visit Mount Sinai. This mountain, renowned in the Abrahamic religions—especially Judaism—is believed to be the place where the Israelites spent three months during their 40-year journey through the wilderness. The tourists listened with interest as their guide explained that the Israelites had reached this mountain after leaving Egypt and had set up camp there.

According to traditional accounts of the Abrahamic faiths, God gave His divine law to the Israelites at this very site. One of the tourists asked in surprise, "Is this the place where the Golden Calf of Samiri was made?" The guide smiled and replied, "Yes, exactly. This is also where Yahweh first revealed Himself to Moses and appeared to him in the form of fire within a burning bush."

The tourists listened in awe as the guide explained that God had commanded Moses to remove his sandals, as he was standing on holy ground. The guide added, "This mountain is also known by other names, such as Horeb, Jabal Allah, al-Tur, and Jabal Musa. It is even mentioned in the Qur'an, where God swears by it."

With a deep sense of wonder and reverence, the tourists departed Mount Sinai, carrying with them an experience they would never forget.

Rate

A

Mistral:

Multiple Choice Question: What emotion did the tourists feel at Mount Sinai that made the experience unforgettable for them? A) Fear B) Awe and Respect C) Fatigue D) Disappointment Answer: B	C
True/False Judgment 1: The tourists remembered the experience because they were amazed by Mount Sinai's history. Answer: True	C
True/False Judgment 2: The tourists remembered the experience because of physical exhaustion. Answer: False	A
Cause and Effect: Cause: The tourists learned about the history and sacredness of Mount Sinai. Effect: The tourists felt awe and respect.	A
Cloze: Sentence: The tourists left Mount Sinai with a sense of "wonder" and respect. Answer: wonder	A
Explanation Question: Why did the tourists remember this experience so vividly? Answer: Because they felt awe and respect for the history and sacredness of Mount Sinai.	C
Counterfactual Question: If the tourists had not known about the history of Mount Sinai, would they still have remembered the experience as vividly? Answer: No, because it was the amazement and respect for the site's history and sanctity that made the experience memorable..	C

Chatgpt:

يهودیت (به عبری: יהדות، Yahădūt) دین، فلسفه، فرهنگ و سبک زنگی قوم یهود است. یهودیت، نخستین دین ابراهیمی و نزدیکی به عقیده سیاری، نخستین دین سازمان یافته و یکتایر است. یهودیت شامل مجموعه گستردگی از دیدگاه‌های از دین ابراهیمی، تورات، تورات، شاخه‌ها، و متون مذهبی می‌باشد. امروز اکثر یهودیان پیرو سه شاخه ارتودوکس، لیبرال، و محافظه‌کار و اقلیتی هم پیرو بازسازی‌خواهی، قرائیم و یهودیت اوتمنیستی هستند؛ از لحاظ دیدگاه‌های مذهبی این شاخه‌ها سیار متفاوتند. در حالی که یهودیان ارتودوکس باور به وحی تورات از طرف یهود، پیروی سخت از قوانین مذهبی، و ظهور مسیحیان موعد دارند، یهودیان لیبرال دارای باور به زنگی مدرن، استقلال شخصی، آزاداندیشی هستند. به این یهودیان مذهبی، یهودیت، پیمان یهود با بنی اسرائیل است.

يهودیت کانون پیشتر تون و سنت‌ها است و تاریخ، اصول و اخلاق آن بر ادیان ابراهیمی پس از خود از جمله مسیحیت و اسلام تأثیر بسزایی گذاشته است. بر طبق قوانین یهودیت، چه نوکیشان (افرادی که به واسطه ازدواج در تعداد بسیار محدود در هر نسل وارد جامعه یهودی شده‌اند) و چه افرادی که از مادر و پدری یهودی متولد شده‌اند، یهودی خطاب می‌شوند. گرچه یهودیت یک دین است، ولی یهودیان یک گروه قومی-مذهبی باحساب می‌آیند. مسئله نژادی و نسبت‌های دی‌ان‌ای فراوان این مسئله را تأثیر می‌کند. دین یهودیت دارای تاریخ پیوسته سه تا چهار هزار ساله است. پیرو (۶۶۷-۶۶۰) و از اهای عبری است، به معنای ستوده؛ همچنین صفتی اکتسابی است که به ساکنان یهودی، قسمی از سرزمین اسرائیل که سبط یهودا بر آن زنگی می‌کرند، داده می‌شود، ولی همچنین برخی معقدنده در گذشته همه بنی اسرائیل همه را نیز به این نام خوانده‌اند. در ایران از عنوان «کلیمی» یعنی پیرو موسی کلیم‌الله (لقب اسلامی موسی) نزد برای نامیدن یهودیان استفاده می‌شود. به این یهودیان مذهبی، اصل یهودیت از وعده خدا با ابراهیم، اسحاق، و یعقوب شروع می‌شود و رسمًا با تحویل قوانین تورات به موسی و ورود بنی اسرائیل به سرزمین موعد کامل می‌شود.

پیروان یهودیت در بسیاری از کشورهای جهان ساکن هستند. جمعیت یهودیان جهان ۱۳۰۷۴۶۱۰۰ نفر (۱۷۹۳۶۴۰۰)، شامل افرادی که بر حسب هلاخا یهودی قلم داده شوند، معاذل حدود ۱۹٪، جمعیت جهان می‌باشد. حدود ۶۱ میلیون نفر ساکن اسرائیل و ۵۲۷۵۰۰۰ (۴۸۰ میلیون کل) نفر ساکن ایالات متحده آمریکا و ۱۵ میلیون نفر در اروپا، و مابقی در آقیانوسیه، قاره آمریکا و آفریقا ساکنند. یهودیان ایران دارای تاریخ ۲۵۰۰ ساله هستند و بر حسب امارت ۱۳۹۱، حدود ۸۷۵۶ یهودی ساکن ایران هستند.

Rate
سارا که در دانشگاه تهران در رشته ادبیات تطبیقی تحصیل می‌کرد، برای پژوهشی پایان‌نامه‌اش تصمیم گرفت درباره یهودیت تحقیق کند. با وجود اینکه خوشناس مسلمان بود، علاقمند بود که پیشتر درباره فرهنگ‌ها و ادیان مختلف بداند. یک روز از طریق یکی از استادیش با خانواده‌ای یهودی در تهران آشنا شد. آنها بازی داری باز از سارا دعوت کردند تا در یکی از مراسم‌های شبات آنان شرکت کند.

در شب جمعه، وقتی سارا وارد خانه شد، فضای متفاوتی را حس کرد. خانواده با احترام خاصی شمع‌های شبات را روشن کردند و بخش‌هایی از تورات را با صدای آرام می‌خواندند. سارا از مازیار، پسر بزرگ خانواده، درباره دیدگاه‌های مختلف در بین یهودیان پرسید. او توضیح داد که یهودیان انواع دیدگاه‌های دینی دارند: برخی مانند اقوام‌شان در اصفهان، باورهای محافظه‌کارانه دارند و به قوانین مذهبی سخت‌گیرانه پایبندند؛ برخی دیگر در اروپا پیشتر تمایلات لیبرال دارند و آزادی تفکر را ارج می‌نهند.
--

مازیار همچنین توضیح داد که بر اساس شریعت یهودیت، کسی که از مادر یهودی به دنیا آمده، یهودی محسوب می‌شود، اما در برخی موارد هم افرادی با ازدواج و تغییر دین به این جامعه وارد می‌شوند. او به سارا گفت که دین یهودیت قسمی نزدیک به چهار هزار سال دارد و تأثیر فراوانی بر ادیان ابراهیمی دیگر گذاشته است. سارا با شنیدن داستان عهد خدا با ابراهیم تا موسی، پیشتر تحت تأثیر این سنت عمیق فرار گرفت.	A
--	---

وقتی بیرون آمد، شب بارانی بود اما ذهنش روشن و پر از سوال‌های تازه بود. از خودش پرسید که چطور یک قوم با این پیشنهاد، می‌تواند هم یکتایرستی باستانی را حفظ کند و هم در کشورهای مختلف با فرهنگ‌های گوناگون سازگار شود.

Chatgpt:

Multiple Choice Question: چرا احتمال دارد سارا دعوت خانواده یهودی برای شرکت در مراسم شبایت را پنیرفت؟ Options: A) چون می خواست برای شب جمیعه سرگرمی داشته باشد. B) چون علاقه داشت فرهنگ مقاومتی را از نزدیک تجربه کند. C) چون دنبال تغییر دین بود. D) چون استادش او را مجبور کرد. Answer: B	B
True/False Judgment 1: سارا پس از مراسم شبایت، سوالاتی جدید درباره یهودیت در ذهنش شکل گرفت. Answer (True) (درست)	A
True/False Judgment 2: مازیار توضیح داد که تنها راه یهودی شدن، تولد از پدر یهودی است. Answer: (False) (نادرست)	A
Cause and Effect: Cause: سارا در مراسم شبایت شرکت کرد و با خانواده یهودی گفتگو کرد. Effect: ذهنش برای پرسیدن سوالات جدید درباره یهودیت و فرهنگ دینی بازتر شد.	A
Cloze: Sentence: با وجود اینکه خودش مسلمان بود، سارا علاقمند بود که بیشتر درباره «_____» بداند. Answer: فرهنگ‌ها و ادیان مختلف	A
Explanation: Question: چرا مازیار درباره دیدگاه‌های مختلف یهودیان در نقاط مختلف جهان صحبت کرد؟ Answer: او می خواست به سارا نشان دهد که یهودیت دیدگاه‌های متنوع دارد و نمی‌توان آن را فقط با یک نگرش شناخت.	A
Counterfactual: Question: اگر سارا در آن شب بارانی به مراسم شبایت نرفته بود، چه اتفاقی می‌افتد؟ Answer: احتمالاً درک شخصی و مستقیم کمتری از سنت‌های یهودی پیدا می‌کرد و سوالات ذهنی‌اش نیز بوجود نمی‌آمد.	A

Chatgpt:

(Judaism) is the religion, philosophy, culture, and way of life of the Jewish people. It is considered the first Abrahamic religion and, by many scholars, the first organized monotheistic faith. The Hebrew Bible, known as the Tanakh, which includes the Torah, Nevi'im (Prophets), and Ketuvim (Writings), constitutes the sacred texts of Judaism. Judaism encompasses a wide range of religious views, rituals, denominations, and religious literature. Today, most Jews identify with one of three main branches: Orthodox, Liberal (Reform), and Conservative. There are also smaller groups such as Reconstructionist Judaism, Karaite Judaism, and Humanistic Judaism. These branches differ significantly in terms of religious beliefs and practices. Orthodox Jews strictly follow traditional religious laws and believe the Torah was divinely revealed by Yahweh and that a Messiah will one day arrive. In contrast, Liberal Jews promote modernity, personal freedom, and individual interpretation of religious life. Religious Jews believe that Judaism represents a covenant between Yahweh and the children of Israel. Judaism serves as the foundation of many texts, traditions, and ethical principles, and has profoundly influenced later Abrahamic religions such as Christianity and Islam. According to Jewish law (Halakha), a person is considered Jewish if they are born to a Jewish mother or have formally converted to Judaism. Although Judaism is a religion, Jews are also seen as an ethno-religious group, with DNA studies and historical records supporting this meaning "praised" in (יהודה). Judaism has a continuous history stretching back 3,000 to 4,000 years. The word *Yehudah* Hebrew, originally referred to the tribe of Judah and later to the inhabitants of the kingdom of Judea. Eventually, it became a term used for all Israelites. In Iran, Jews are also called Kalimi, referring to Moses (Kalim-ullah), the prophet of Judaism in Islamic tradition. Religious Jews believe Judaism begins with God's covenant with Abraham, Isaac, and Jacob, and is officially established when the Torah is given to Moses and the Israelites enter the Promised Land. Jewish communities exist all over the world. The global Jewish population is approximately 13.7 million (or up to 17.9 million when including individuals not recognized as Jewish by Halakha), which constitutes about 0.19% of the world population. Around 6.1 million live in Israel, 5.3 million in the United States, 1.5 million in Europe, and the rest are spread across Oceania, the Americas, and Africa. Iran's Jewish community, which dates back over 2,500 years, currently numbers around 8,756, based on the 2012 census.

Sara, a student of Comparative Religions at the University of Tehran, chose to research Judaism for her thesis project. Although she was Muslim, she was deeply interested in learning more about different cultures and faiths. One day, through one of her professors, she was introduced to a Jewish family living in Tehran. They warmly invited her to attend one of their Shabbat ceremonies. On Friday evening, as Sara entered their home, she immediately sensed a different atmosphere. The family respectfully lit the Shabbat candles, and softly recited passages from the Torah. Curious, Sara asked Maziyar, the eldest son, about the diversity of beliefs among Jews. He explained that Judaism includes a wide range of religious perspectives: some, like their relatives in Isfahan, are more conservative and strictly follow religious laws, while others, particularly in Europe, adopt more liberal views and value intellectual freedom.	Rate
Maziyar also shared that, according to Jewish law (Halakha), a person born to a Jewish mother is considered Jewish, though in some cases, individuals may join the community through marriage or conversion. He told her that Judaism has a history of nearly four thousand years, and it has had a profound impact on other Abrahamic religions. Listening to the story of God's covenant with Abraham through to Moses, Sara felt deeply moved by the rich and enduring tradition. When she left the house, the night was rainy, but her mind felt clear and filled with new questions. She asked herself how a people with such ancient roots could preserve monotheism while also adapting to so many different cultures across the world.	A

Chatgpt:

Multiple Choice Question: Why did Sara likely accept the invitation from the Jewish family to attend the Shabbat ceremony? Options: A) Because she wanted entertainment on Friday night. B) Because she was interested in experiencing a different culture up close. C) Because she was seeking to change her religion. D) Because her professor forced her. Answer: B	A
True/False Judgment 1: After the Shabbat ceremony, Sara had new questions about Judaism in her mind. Answer: True	A
True/False Judgment 2: Maziyar explained that the only way to become Jewish is to be born of a Jewish father. Answer: False	A
Cause and Effect: Cause: Sara attended the Shabbat ceremony and spoke with the Jewish family. Effect: Her mind became more open to asking new questions about Judaism and religious culture.	A
Cloze: Sentence: Although she was Muslim herself, Sara was interested in learning more about " ____ ". Answer: different cultures and religions	A
Explanation: Question: Why did Maziyar talk about the different viewpoints of Jews in various parts of the world? Answer: He wanted to show Sara that Judaism has diverse perspectives and cannot be understood through a single viewpoint.	A
Counterfactual: Question: What would have happened if Sara had not gone to the Shabbat ceremony on that rainy night? Answer: She probably would have had less personal and direct understanding of Jewish traditions, and her mental questions would not have arisen.	A

		Agreement %	κ (Quadratic)
Scenario		85.6	0.405
Task 1	Multiple-choice questions	94.4	0.201
Task 2	True/False statement	97.8	0.409
Task 3	Explanation	98.0	0.493
Task 4	Cloze	96.8	0.373
Task 5	Cause-effect inference	96.0	0.282
Task 6	Counterfactual reasoning	93.8	0.240

Table 10: Inter-annotator agreement across scenario outputs and task types.

E Inter-Annotator Agreements

As shown in Table 10, although the overall percent agreement between annotators is high (e.g., 85.6% at the scenario level), Cohen’s κ remains moderate (e.g., 0.405). This discrepancy arises because the annotations are heavily skewed toward one dominant class—typically rating “A”. Since classes “B” and “C” occur rarely in the dataset, the expected agreement by chance (p_e) becomes disproportionately high, which mathematically reduces the value of κ . This is a well-known limitation of κ in imbalanced rating distributions, where even near-perfect agreement on the majority class can result in deceptively low scores. In our case, the high agreement primarily reflects consistent use of the dominant rating, while κ penalizes the lack of variability across categories.

F Representative Gemini-2.5-Flash Outputs and Their Human-Annotated Revisions

Below, we present three examples generated by the Gemini-2.5-Flash model. In the first two cases, the model used inappropriate metaphors (e.g., comparing the human body to a garden) that were not aligned with the topic, resulting in scenarios that lacked clarity and semantic coherence. As a result, human annotators regenerated the scenarios using the model to produce more accurate and meaningful versions. In contrast, the third example was well-formed from the start — the initial generation by the model was both coherent and natural, requiring no modifications and receiving a high-quality rating from annotators. Each example includes the original Wikipedia passage, the scenario generated from it, and a set of six tasks constructed based on the scenario.

Gemini:

سی (به انگلیسی: C،) یک زبان برنامه‌نویسی از نوع همانمنظره، کامپایل‌شونده، سطح میانی، ساخت‌یافته، دستوری و روندگار می‌باشد که در سال ۱۹۷۲ توسط دنیس ریچی در آزمایشگاه‌های بل ساخته شد. در سال ۱۹۶۷ مارتین ریچاردز زبان BCPL را برای نوشتن نرم‌افزارهای سیستم‌عامل و کامپایلر در دانشگاه کمبریج ابداع کرد. سپس در سال ۱۹۷۰ کن تامسون زبان B را بر مبنای ویژگی‌های زبان BCPL نوشت و از آن برای ایجاد اولین نسخه‌های سیستم‌عامل یونیکس در آزمایشگاه‌های بل استفاده کرد. زبان C در سال ۱۹۷۲ توسط دنیس ریچی از روی زبان B و BCPL در آزمایشگاه بل ساخته شد و ویژگی‌های جدیدی همچون نظارت بر نوع داده‌ها نیز به آن اضافه شد. ریچی از این زبان برای ایجاد سیستم‌عامل یونیکس استفاده کرد اما بعداً اکثر سیستم‌عامل‌های دیگر نیز با همین زبان نوشته شدند. این زبان با سرعت بسیاری گسترش یافت و چاپ کتاب "The C Programming Language" در سال ۱۹۷۸ توسط برایان کرینیگان و ریچی باعث رشد روزافزون این زبان در جهان شد.	Rate
استفاده گسترده این زبان در انواع رایانه‌ها و سخت‌افزارهای مختلف باعث شد که نسخه‌های مختلفی از این زبان باوجود آید که با یکی‌گر ناسازگار بودند. در سال ۱۹۸۳ مؤسسه استانداردهای ملی امریکا (ANSI) کمیته‌ای موسوم به X3J11 را مأمور کرد تا یک تعریف فاقد ابهام و مستقل از ماشین را از این زبان تدوین نماید. در سال ۱۹۸۹ این استاندارد تحت عنوان C ANSI به تصویب رسید و سپس در سال ۱۹۹۰، سازمان بین‌المللی استانداردسازی (ISO) نیز این استاندارد را پذیرفت و مستندات مشترک آنها تحت عنوان C منتشر گردید.	
در دنیای برنامه‌نویسی، روزی بود که "بافنده کد" (یک زبان برنامه‌نویسی کامپایل‌شونده، همانمنظره و سطح میانی) ظهر کرد. این بافنده کد که ساخت‌یافته و دستوری بود و روش روندگرا داشت، در سال ۱۹۷۲ توسط اقای دنیس ریچی در کارگاه‌های "بل" ابداع شد. داستان از این قرار بود که پیش از آن، در سال ۱۹۶۷، اقای مارتین ریچاردز "پروتوتایپ ساز" (BCPL) را در اکادمی کمبریج برای نوشتن نرم‌افزارهای سیستم‌عامل و کامپایلرها معرفی کرده بود. بیسی، در سال ۱۹۷۰، اقای کن تامسون، با الهام از "پروتوتایپ ساز"، "سازنده نسخه دوم" (زبان B) را نوشت و از آن در همان کارگاه‌های بل برای خلق اولین سخنهای "سیستم‌عامل جهانی" (یونیکس) بهره گرفت.	
اما "بافنده کد" گام بلندتری برداشت؛ ریچی آن را بر پایه "سازنده نسخه دوم" و "پروتوتایپ ساز" بنایه داد و ویژگی‌های جدیدی چون نظارت دقیق بر نوع داده‌ها را به آن افزود. ریچی از "بافنده کد" برای توسعه همین "سیستم‌عامل جهانی" استفاده کرد و سپس، این ایزار افتخار قدرتمند شد که اکثر سیستم‌عامل‌های دیگر نیز با آن نوشته شدند. رشد "بافنده کد" سراسام‌اور بود، خصوصاً پس از انتشار "کتاب راهنمای بافنده کد" در سال ۱۹۷۸ توسط اقایان برایان کرینیگان و ریچی که شهرت جهانی آن را تضمین کرد.	
با این حال، استفاده گسترده از این ایزار در کامپیوترها و سخت‌افزارهای مختلف باعث شد "گویش‌های" ناسازگاری از آن پیدا آید. در سال ۱۹۸۳، "نگهبانان استاندارد ملی امریکا" (ANSI) گروه ویژه‌ای به نام X3J11 تشکیل دادند تا تعریفی بدون ابهام و مستقل از ماشین برای "بافنده کد" ارائه دهند. سرانجام، در سال ۱۹۸۹، این استاندارد با عنوان "بافنده کد ANSI" تایید شد و در سال ۱۹۹۰، "اتحادیه بین‌المللی هماهنگی" (ISO) نیز آن را پذیرفت و "طرح مشترک" نهایی با عنوان "بافنده کد ANSI/ISO" منتشر گردید تا همه از یک زبان واحد سخن بگویند	C

Gemini:

<p>Multiple Choice Question:</p> <p>تصمیم گرفتند یک گروه ویژه برای بافته کد (ANSI) چرا نگهبانان استاندارد ملی امریکا نشکل دهند؟</p> <p>A) برای رقابت با «اتحادیه بین‌المللی هماهنگی (ISO).</p> <p>B) برای حل مشکل ناسازگاری در نسخه‌های مختلف زبان.</p> <p>C) به دلیل پیچیدگی بیش از حد زبان برای برنامه‌نویسان متندی.</p> <p>D) برای معرفی قابلیت‌های جدید و پیشرفته به زبان.</p>	<p>Answer: B</p>	B
<p>True/False Judgment 1:</p> <p>قبل از ابداع «بافته کد»، تلاش‌هایی برای توسعه زبان‌های برنامه‌نویسی مشابه وجود داشت که پایه‌ای برای آن شدند.</p> <p>Answer:</p> <p>(درست) (True)</p>		B
<p>True/False Judgment 2:</p> <p>انتشار «کتاب راهنمای بافته کد» پیش از اختراع خود زبان «بافته کد» اتفاق افتاد و به معرفی اولیه آن کمک کرد.</p> <p>Answer:</p> <p>(نادرست) (False)</p>		C
<p>Cause and Effect:</p> <p>Cause:</p> <p>بافته کد به اندازه‌ای قدرتمند و کارآمد بود که بتوان با آن سیستم‌عامل‌های مختلف را توسعه داد.</p> <p>Effect:</p> <p>این زبان به سرعت مورد استفاده گسترده‌ای قرار گرفت و شهرت جهانی پیدا کرد.</p>		B
<p>Cloze:</p> <p>Sentence:</p> <p>هدف نهایی انتشار استاندارد «بافته کد» ANSI/ISO این بود که همه بتوانند.</p> <p>Answer:</p> <p>بدون مشکل از زبان استفاده کنند و برنامه‌های سازگار بنویسند</p>		B
<p>Explanation:</p> <p>Question:</p> <p>چرا اضافه شدن ویژگی «نظرارت دقیق بر نوع داده‌ها» به «بافته کد» یک گام بلند به حساب می‌آمد؟</p> <p>Answer:</p> <p>این ویژگی به کاهش خطاهای افزایش پایداری برنامه‌ها کمک می‌کرد، زیرا نوع داده‌های ورودی و خروجی را بررسی می‌کرد و از عملیات ناسازگار جلوگیری می‌نمود.</p>		B
<p>Counterfactual:</p> <p>Question:</p> <p>اگر نگهبانان استاندارد ملی امریکا (ANSI) و اتحادیه بین‌المللی هماهنگی (ISO) نتوانسته بودند استانداردی واحد برای «بافته کد» ارائه دهند، چه اتفاقی می‌افتد؟</p> <p>Answer:</p> <p>استفاده از «بافته کد» بسیار دشوارتر می‌شد و برنامه‌های نوشته شده با گوشی‌های مختلف به راحتی قابل همکاری یا جابجایی بین سیستم‌های مختلف نیزند.</p>		C

Gemini:

<p>C (in English: C) is a general-purpose, compiled, mid-level, structured, imperative, and procedural programming language, developed in 1972 by Dennis Ritchie at Bell Labs. In 1967, Martin Richards developed the BCPL language at the University of Cambridge to write operating system software and compilers. Then in 1970, Ken Thompson created the B language based on features from BCPL, and used it to develop the first versions of the Unix operating system at Bell Labs. The C language was created in 1972 by Dennis Ritchie at Bell Labs, evolving from the B and BCPL languages. It introduced new features such as type checking. Ritchie used C to develop the Unix operating system, and later, most other operating systems were also written in C. C spread rapidly, and the publication of the book "The C Programming Language" in 1978 by Brian Kernighan and Dennis Ritchie significantly boosted its popularity worldwide. Due to its widespread use across various computers and hardware platforms, multiple versions of the language emerged, which were often incompatible with one another. In 1983, the American National Standards Institute (ANSI) formed a committee called X3J11 to create an unambiguous, machine-independent definition of the language. In 1989, this standard was approved as ANSI C, and in 1990, the International Organization for Standardization (ISO) also adopted it. The combined documentation was published under the name ANSI/ISO C.</p>	Rate
<p>C is a general-purpose, compiled, mid-level, structured, procedural programming language created in 1972 by Dennis Ritchie at Bell Laboratories. In 1967, Martin Richards invented the BCPL language to write operating system software and compilers at the University of Cambridge. Then, in 1970, Ken Thompson wrote the B language based on the features of BCPL and used it to develop the first versions of the Unix operating system at Bell Labs. In 1972, Dennis Ritchie created the C language at Bell Labs, building on the B and BCPL languages, and added new features such as data type checking. Ritchie used this language to develop the Unix operating system, but later most other operating systems were also written in C. The language spread rapidly, and the publication of the book <i>The C Programming Language</i> in 1978 by Brian Kernighan and Ritchie contributed to its growing popularity worldwide. The widespread use of C on various computers and hardware led to the emergence of incompatible versions of the language. In 1983, the American National Standards Institute (ANSI) formed a committee called X3J11 to develop an unambiguous, machine-independent definition of the language. In 1989, this standard was approved as ANSI C, and in 1990, the International Organization for Standardization (ISO) adopted the standard. The joint documents were published under the title ANSI/ISO C, so that everyone could refer to a unified language.</p>	C

Gemini:

<p>Multiple Choice Question: Why did the American National Standards Institute (ANSI) decide to form a special committee for “Code Weaver”?</p> <p>A) To compete with the International Organization for Standardization (ISO). B) To solve the problem of incompatibility between different versions of the language. C) Because the language was too complex for beginner programmers. D) To introduce new and advanced features to the language.</p> <p>Answer: B</p>	B
<p>True/False Judgment 1: Before the invention of “Code Weaver,” there were efforts to develop similar programming languages that became its foundation.</p> <p>Answer: True</p>	B
<p>True/False Judgment 2: The “Code Weaver Handbook” was published before the language itself was invented and helped introduce it.</p> <p>Answer: False</p>	C
<p>Cause and Effect: Cause: “Code Weaver” was powerful and efficient enough to be used for developing operating systems. Effect: The language quickly gained widespread adoption and became globally recognized.</p>	B
<p>Cloze: Sentence: The ultimate goal of publishing the ANSI/ISO “Code Weaver” standard was to ensure that everyone could “_____.”</p> <p>Answer: use the language without issues and write compatible programs</p>	B
<p>Explanation: Question: Why was the addition of “strict type checking” to “Code Weaver” considered a major advancement?</p> <p>Answer: Because it helped reduce bugs and improve program stability by verifying input and output data types and preventing incompatible operations.</p>	B
<p>Counterfactual: Question: What would have happened if ANSI and ISO had failed to agree on a unified standard for “Code Weaver”?</p> <p>Answer: Using “Code Weaver” would have become much more difficult, and programs written in different dialects would not easily work together or be portable across systems.</p>	C

Gemini:

سال‌ها پیش، در فضای پر‌هیاهوی آزمایشگاه‌های بل، دنیس ریچی مشغول خلق چیزی بود که قرار بود دنیای محسابات را دگرگون کند. در سال ۱۹۷۲، او ابزاری نو ساخت: یک زبان برنامه‌نویسی جدید که مثل یک جعبه‌ای از هم‌منظوره بود و می‌شد از آن برای کارهای سیاری استفاده کرد. این ابزار، پیش از به کار رفتن، نیاز به ترجمه و آماده‌سازی داشت، شبیه به نقشه‌هایی که باید به کمایشین تبدیل شوند تا کامپیوتر بفهمشان. این زبان در عین اینکه به جزئیات ماشین‌ها توجه داشت، از چارچوب‌های مشخصی پیروی می‌کرد، کامپیکام دستورات را اجرا می‌کرد و روش‌های مشخصی داشت.	Rate
ریچی ایده‌اش را از کارهای قبلی الهام گرفته بود. در سال ۱۹۶۷، مارتن ریچاردز در دانشگاه کمبریج، سنت‌گینیز زبان BCPL را برای نوشتن نرم‌افزارهای سیستم‌عامل و کامپایلرها گذاشته بود. بعدتر، در سال ۱۹۷۰، کن تامسون با الهام از ویژگی‌های BCPL، زبان B را نوشت و آن را در همان آزمایشگاه‌های بل برای خلق اولین نسخه‌های سیستم‌عامل بونیکس به کار برد.	A
ریچی، زبان C را بر اساس همین BCPL و B بنانهاد و ویژگی‌های جدیدی مثل نظرارت دقیق بر نوع داده‌ها را به آن افزود. او این زبان جدید خود برای ساخت سیستم‌عامل بونیکس استفاده کرد و طولی نکشید که بسیاری از سیستم‌عامل‌های دیگر نیز با همین ابزار قادرند نوشته شدند.	
"The C" کتاب، در سال ۱۹۷۸ مانند موجی سهمگین در دنیا پیچید. در سال ۱۹۷۸، چاپ کتاب	
"Programming Language" که حاصل همکاری ریچی و برایان کرنیگان بود، به این محبوبیت جهانی دامن زد. اما استفاده بی‌حد و حصر این زبان در انواع رایانه‌ها و سخت‌افزارهای گوتانگون، کمک به تفاوت‌ها و ناسازگاری‌هایی منجر شد. برای حل این مشکل، در سال ۱۹۸۳، مؤسسه استانداردهای ملی آمریکا (ANSI)، گروه ویژه‌ای به نام X3J11 را مسئول تعریف یک استاندارد جامع و روش برای C کرد، استانداردی که فارغ از نوع ماشین باشد. سرانجام در سال ۱۹۸۹، این استاندارد تحت عنوان ANSI C به تصویب رسید و یک سال بعد، در سال ۱۹۹۰، سازمان بین‌المللی استانداردسازی (ISO) نیز آن را پذیرفت و مستاندار مشترک آن‌ها بنام ANSI/ISO C منتشر شد. اینکوئه بود که C جایگاه خود را به عنوان یکی از ستون‌های اصلی دنیای برنامه‌نویسی تثبیت کرد.	

Gemini:

Multiple Choice Question:	Rate
<p>به سرعت به یکی از زبان‌های برنامه‌نویسی پرکاربرد تبدیل شد؟ C چرا زبان به دلیل سادگی مطلق آن که یادگیری آش را برای همگان آسان می‌کرد قابلیت استفاده آن برای ایجاد سیستم‌عامل‌ها و نرم‌افزارهای اساسی متنوع انتشار زودهنگام و گسترش دنیاگیر را اهتمای آن پیش از هر کاربرد عملی چون اولین زبان برنامه‌نویسی سطح بالا بود که توسط کامپیوترها فهیمده می‌شد</p> <p>Answer: B</p>	A
<p>True/False Judgment 1: عده‌تاً به دلیل نیاز به حل مشکلات ناسازگاری در کاربردهای گسترده C استاندارد سازی زبان آن صورت گرفت.</p> <p>Answer: (True)</p>	A
<p>True/False Judgment 2: بدون نیاز به هیچ‌گونه آمده‌سازی یا ترجمه مستقیم توسط کامپیوترها قابل فهم بود C زبان</p> <p>Answer: (False)</p>	A
<p>Cause and Effect: Cause: انتشار یک کتاب مرجع جامع و استاندارد برای یک زبان برنامه‌نویسی جدید Effect: تسريع و تسهیل پنیرش و گسترش آن زبان در میان برنامه‌نویسان</p>	A
<p>Cloze: Sentence: قبل از اینکه توسط کامپیوترها اجرا شود، مانند نقشه‌های معماری است که باید به زبان _____ « تبدیل گردد تا قابل فهم و ساخت باشند».</p> <p>Answer: کد اجرایی</p>	A
<p>Explanation: Question: کمک C به شهرت جهانی زبان "The C Programming Language" چگونه انتشار کتاب کرد؟</p> <p>Answer: شد و C این کتاب با ارانه یک مرجع جامع و معتبر، باعث استاندارد سازی دانش مربوط به زبان دسترسی برنامه‌نویسان در سراسر جهان به آن را تسهیل کرد</p>	A
<p>Counterfactual: Question: اگر مؤسسه ANSI برای زبان C استانداردی تعریف نمی‌کرد، چه اتفاقی ممکن بود بیفتند؟</p> <p>Answer: بدون یک استاندارد یکپارچه، ناسازگاری‌ها بین نسخه‌های مختلف C در سخت‌افزارهای گوناگون ادامه می‌یافتد و این امر می‌توانست مانع از گسترش و تثبیت موقعیت آن به عنوان یک زبان برنامه‌نویسی بنیادین شود.</p>	A

Gemini:

Rate
A

Years ago, in the bustling halls of Bell Labs, Dennis Ritchie was busy creating something that would revolutionize the world of computing. In 1972, he developed a new tool—a programming language that acted like a versatile toolbox, capable of handling a wide range of tasks. Before it could be used, this tool needed to be translated and prepared, much like architectural plans that must be turned into machine code for a computer to understand. While this language was mindful of machine-level details, it also followed structured frameworks, executed instructions step by step, and adhered to well-defined procedures.

Ritchie's idea was inspired by earlier work. In 1967, Martin Richards at the University of Cambridge laid the foundation for the BCPL language, which was designed for writing system software and compilers. Later, in 1970, Ken Thompson, drawing on the features of BCPL, created the B language and used it at Bell Labs to build the early versions of the UNIX operating system. Ritchie based the C language on both BCPL and B, enhancing it with new features like strict data type checking. He then used C to build the UNIX operating system itself—and before long, many other operating systems were written using this powerful language.

The fame of C spread like a tidal wave across the world. In 1978, the publication of *The C Programming Language*, co-authored by Ritchie and Brian Kernighan, further accelerated its global popularity. However, the widespread use of C across different computers and hardware eventually led to incompatibilities and variations. To address this, in 1983, the American National Standards Institute (ANSI) formed a special committee named X3J11 to define a clear, machine-independent standard for C. This effort resulted in the adoption of ANSI C in 1989, and a year later, in 1990, the International Organization for Standardization (ISO) also approved it. The joint documentation, published as ANSI/ISO C, ensured that programmers around the world could speak the same language. In this way, C firmly established itself as one of the fundamental pillars of the programming world.

Gemini:

<p>Multiple Choice Question: Why did the C programming language quickly become one of the most widely used languages?</p> <p>A) Because it was extremely simple and easy for everyone to learn. B) Because it could be used to develop operating systems and essential software. C) Because its handbook was published widely before any real use. D) Because it was the first high-level language computers could understand directly.</p> <p>Answer: B</p>	A
<p>True/False Judgment 1: The standardization of the C language was primarily done to resolve incompatibilities caused by its widespread use.</p> <p>Answer: True</p>	A
<p>True/False Judgment 2: C could be directly understood by computers without any preparation or translation.</p> <p>Answer: False</p>	A
<p>Cause and Effect Cause: The release of a comprehensive and standardized reference book for a new programming language. Effect: Accelerated and facilitated the language's acceptance and growth among programmers.</p> <p>Answer: Causal</p>	A
<p>Cloze: Sentence: Before being executed by computers, the C language is like architectural blueprints that must be converted into "____" to be understood and built.</p> <p>Answer: executable code</p>	A
<p>Explanation: Question: How did the publication of <i>The C Programming Language</i> contribute to the global popularity of C?</p> <p>Answer: The book provided a comprehensive and reliable reference that helped standardize knowledge of C and made it more accessible to programmers worldwide.</p>	A
<p>Counterfactual: Question: What might have happened if ANSI had not defined a standard for the C language?</p> <p>Answer: Without a unified standard, incompatibilities across different versions of C would likely have persisted, preventing the language from becoming a foundational tool in programming.</p>	A

Gemini:

<p>در بدن انسان و گروهی دیگر از جاتوران، مثانه ادرار را از طریق لوله‌ای به نام میزراه، پیشابراد، گمیزراه، مجرای ادراری، تیبل وی یا اورترا (به انگلیسی: Urethra) تخلیه می‌کند. میزراه در زنان کوتاه است، اما در مردان این مجرا درازتر است و از میان غده پروستات و آلت تناسلی می‌گذرد.</p> <p>اسفکتر خارجی مثانه که مانع دفع غیرارادی ادرار می‌شود در ایندای میزراه قرار دارد. چون میزراه در زنان کوتاهتر است و به همین دلیل عفونت در دستگاه ادراری-تناسلی زنان چندین برابر بیشتر از مردان است.</p> <p>اندازه میزراه در مردان بین ۲۰ تا ۲۵ سانتی‌متر بوده و در سر آلت تناسلی باز می‌شود و اندازه میزراه در زنان نیز بین ۵/۲ تا ۵ سانتی‌متر است.</p>	
<p>در یک باغ بزرگ، دو مخزن آب حیاتی وجود داشت که هر دو وظیفه تخلیه آب مازاد را بر عهده داشتند. هر مخزن، لوله‌ای مخصوص به خود داشت که آب را به بیرون هدایت می‌کرد. این لوله‌ها حکم مجرایی برای خروج آب بودند و در انتهایشان یک شیر کنترل‌کننده دقيق قرار گرفته بود که باز و بسته شدنش کاملاً ارادی بود و مانع هدر رفتن ناخواسته آب می‌شد.</p> <p>لوله مخزن «مردانه» بسیار بلندتر از لوله مخزن «زنانه» بود. اگر دقیق‌تر بگوییم، لوله مردانه تقریباً به اندازه یک خلکش استاندار، یعنی حدود ۲۰ تا ۲۵ سانتی‌متر طول داشت و درست در نوک یک برآمدگی خاص از بدن مخزن بیرون می‌آمد. جالب اینکه این لوله در مسیر خود از میان یک «گلوبگ» مه (غده پروستات) و سپس از دل یک «برج» مرکزی (آلت تناسلی) عبور می‌کرد تا به انتهای خود برسد.</p> <p>اما لوله مخزن «زنانه» بسیار کوتاه‌تر بود، چیزی در حد ۲.۵ تا ۵ سانتی‌متر، یعنی تنها به اندازه یک سرکلید کوچک. همین تفاوت طول باعث می‌شد که نظافت و محافظت از لوله زنانه کمی چالش‌برانگیزتر باشد. راه ورود میکروب‌ها و الودگی‌ها به این لوله کوتاهتر، بسیار مستقیم‌تر و سریع‌تر بود، به همین دلیل احتمال گرفتگی یا الودگی‌های داخلی در سیستم زنانه، چندین برابر بیشتر از سیستم مردانه بود. این تفاوت در طراحی، نگهداری از هر دو سیستم را به روش‌های خاص خود می‌طلبید.</p>	<p>Rate</p> <p>C</p>

Gemini:

<p>Multiple Choice Question:</p> <p>با توجه به توضیحات، کدام یک از دلایل زیر به افزایش احتمال گرفتگی یا الودگی‌های داخلی در سیستم زنانه اشاره دارد؟</p> <p>A) طول زیاد لوله آن باعث جمع‌آوری بیشتر الودگی می‌شود. B) کوتاهی و مستقیم بودن مسیر لوله، ورود عوامل بیماری‌زا را آسان می‌کند. C) عدم وجود شیر کنترل‌کننده در انتهای لوله آن. D) قرار گرفتن لوله در میان یک گلوگاه و برج مرکزی</p> <p>Answer: B</p>	B
<p>True/False Judgment 1:</p> <p>تفاوت در طراحی لوله‌های مخازن، ایجاب می‌کند که برای هر کدام رویکردهای مراقبتی و پداشتی مقاومتی در نظر گرفته شود.</p> <p>Answer: درست (True)</p>	B
<p>True/False Judgment 2:</p> <p>لوله مخزن «مردانه» به دلیل طولانی‌تر بودن مسیرش، بیشتر از لوله «زنانه» در معرض ورود مستقیم میکروب‌ها و الودگی‌ها قرار دارد.</p> <p>Answer: نادرست (False)</p>	B
<p>Cause and Effect</p> <p>Cause: کوتاهی و مستقیم بودن مسیر لوله مخزن زنانه</p> <p>Effect: افزایش چشمگیر احتمال ورود مستقیم میکروب‌ها و الودگی‌ها به سیستم</p> <p>Close: Sentence: با توجه به تفاوت در طراحی دو سیستم، نگهداری از هر کدام از آن‌ها به «_____» خود را می‌طلبد.</p> <p>Answer: روش‌های خاص</p>	B
<p>Explanation:</p> <p>Question: چرا لوله مخزن «مردانه» با وجود مسیر طولانی‌تر و عبور از گلوگاه و برج مرکزی، کمتر در معرض الودگی‌های داخلی قرار می‌گیرد؟</p> <p>Answer: طول بیشتر این لوله مسیر ورود میکروب‌ها را دشوارتر و طولانی‌تر می‌کند و موانع موجود در مسیر نیز به عنوان فیلتر یا سدی طبیعی عمل می‌کنند.</p>	B
<p>Counterfactual:</p> <p>Question: اگر طول لوله مخزن «زنانه» مانند لوله مخزن «مردانه»، سیار بلندتر بود، چه تغییری در میزان آسیب‌پذیری آن در برابر الودگی ایجاد می‌شد؟</p> <p>Answer: احتمال ورود مستقیم میکروب‌ها و الودگی‌ها به آن به طور قابل توجهی کاهش می‌یافتد و در نتیجه آسیب‌پذیری کمتری داشت.</p>	B

Gemini:

In the human body and some other animals, the urinary bladder expels urine through a tube called the urethra (also known as the urethral canal, urinary duct, or *Utricle Way* in some texts). The urethra is shorter in women, while in men, it is longer and passes through the prostate gland and the penis.

The external urethral sphincter, which prevents involuntary urination, is located at the beginning of the urethra. Because the female urethra is shorter, urinary tract infections (UTIs) are several times more common in women than in men.

In men, the length of the urethra ranges between 20 to 25 centimeters, and it opens at the tip of the penis.

In a large garden, there were two vital water tanks, each responsible for draining excess water. Each tank had its own dedicated pipe that directed the water outward. These pipes acted as channels for water discharge, and at their end was a precisely controlled valve that could be voluntarily opened or closed, preventing unintended water loss. The “male” tank’s pipe was significantly longer than that of the “female” tank. To be more precise, the male pipe measured approximately the length of a standard ruler — about 20 to 25 centimeters — and exited from the tip of a distinct protrusion on the body of the tank. Interestingly, this pipe passed through a critical bottleneck (the prostate gland) and then through a central tower (the penis) before reaching its outlet. In contrast, the female tank’s pipe was much shorter, only around 2.5 to 5 centimeters — about the size of a small key head. This difference in length made hygiene and protection of the female pipe more challenging. The shorter and more direct path allowed germs and contaminants to enter more easily and quickly, which is why the risk of internal blockage or infection in the female system was significantly higher than in the male system. This structural difference meant that each system required its own specific maintenance approach.

Rate

C

Gemini:

<p>Multiple Choice Question: According to the description, which of the following explains the increased risk of blockages or internal infections in the female system? A) Its long tube causes more accumulation of contaminants. B) The short and straight path of the tube makes it easier for pathogens to enter. C) The absence of a control valve at the end of the tube. D) The tube is located between a bottleneck and a central tower. Answer: B</p>	B
<p>True/False Judgment 1: The difference in the design of tank tubes requires that each system be maintained with a different hygiene and care approach. Answer: True</p>	B
<p>True/False Judgment 2: The tube of the “male” tank, because of its longer path, is more exposed to direct entry of germs and contaminants than the “female” one. Answer: False</p>	B
<p>Cause and Effect Cause: The short and straight path of the “female” tank tube Effect: Significantly increases the risk of direct entry of germs and contaminants into the system</p>	B
<p>Cloze: Sentence: Due to the design differences between the two systems, maintaining each of them required its own “_____.” Answer: specific methods</p>	B
<p>Explanation: Question: Why is the “male” tank tube, despite being longer and passing through a bottleneck and central tower, less prone to internal contamination? Answer: Its longer path makes it harder and slower for germs to enter, and the obstacles in the path act as natural filters or barriers.</p>	B
<p>Counterfactual: Question: If the tube of the “female” tank were as long as that of the “male” tank, what change would occur in its vulnerability to contamination? Answer: The risk of direct entry of germs and contaminants would significantly decrease, resulting in lower vulnerability.</p>	B

Gemini:

Rate
<p>بدن ما، شگفتی از مهندسی است! فرض کنید مثا، برای مایعات اضافه است. برای تخلیه این عضو، یک لوله ضروری است که به آن "میزراه" می‌گوییم. این لوله، نقش تخلیه‌کننده اصلی را بازی می‌کند.</p> <p>حالا تصور کنید دو نوع لوله‌کشی متقاولت داریم، یکی برای بدن "مریم" و دیگری برای بدن "رضا". در بدن مریم، لوله تخلیه (میزراه) بسیار کوتاه است، تقریباً بین ۲،۵ تا ۵ سانتی‌متر. این نزدیکی به فضای بیرون، متناسبانه باعث می‌شود که گاهی اوقات "مشکلات خارجی" یا عفونت‌ها، چندین برابر راحت‌تر از لوله مردانه، راه خود را پیدا کرده و سبب ناراحتی شوند. در واقع، عفونت در دستگاه ادراری-تناسلی زنان به همین دلیل، چندین برابر بیشتر از مردان است.</p> <p>اما در بدن رضا، وضعیت فرق می‌کند. لوله میزراه او بسیار دراز‌تر است، طولی حدود ۲۰ تا ۲۵ سانتی‌متر دارد. این لوله بیچ و خم‌هایی دارد و از میان یک "محفظه محافظ" (غده پروستات) و سیس از یک "بخش بیرونی" (آلت تناسلی) عبور می‌کند تا در نهایت به سر آن برسد و محتویات را خارج کند.</p> <p>و یک نکته مهم دیگر: درست در ایندای هر دو این لوله‌های تخلیه، یک شیر کنترل هوشمند قرار دارد که "اسفکتر خارجی مثا" نامیده می‌شود. این شیر و طیفه بسیار مهمی دارد؛ مانند یک نگهبان، مانع از دفع ناخواسته و غیرارادی مایعات می‌شود و به ما امکان کنترل کامل را می‌دهد. این طراحی دقیق، تضمین می‌کند که مثا در بدن، در زمان مناسب و به شیوه‌ای کنترل شده تخلیه شود.</p>

Gemini:

Multiple Choice Question:	Rate
<p>اگر میزراه زنانه نیز مانند میزراه مردانه بلندتر و پیچیده‌تر بود، کدام یک از پیامدهای زیر متحمل‌تر بود؟</p> <p>دفعه ایاعات بدن با سرعت پیشتری انجام می‌شود؟</p> <p>بین زنان کمتر در معرض عفونت‌های ادراری قرار می‌گرفت</p> <p>غده پروستات در زنان نیز توسعه می‌یافتد</p> <p>بیاز به کنترل ارادی دفع ادرار از بین میرفت</p> <p>Answer: B</p>	A
<p>True/False Judgment 1:</p> <p>کوتاه بودن میزراه در زنان، یکی از دلایل اصلی افزایش احتمال ورود عامل بیماری‌زا از محیط بیرون به دستگاه ادراری است.</p> <p>Answer:</p> <p>(True)</p>	A
<p>True/False Judgment 2:</p> <p>بر اساس متن، غده پروستات یک عضو کلیدی برای کنترل ارادی خروج ادرار در مردان است.</p> <p>Answer:</p> <p>(False)</p>	A
<p>Cause and Effect:</p> <p>Cause:</p> <p>وجود یک "شیر کنترل هوشمند" به نام اسفنکتر خارجی مثانه</p> <p>Effect:</p> <p>قابلیت انسان برای کنترل ارادی و جلوگیری از دفع ناخواسته مایعات بدن</p>	A
<p>Cloze:</p> <p>Sentence:</p> <p>طراحی مقاومت سیستم دفع مایعات در زنان نسبت به مردان، به خصوص در طول میزراه، متأسفانه می‌تواند باعث شود که عفونت‌ها «_____»</p> <p>Answer:</p> <p>چندین برابر راحت‌تر راه خود را پیدا کند</p>	A
<p>Explanation:</p> <p>Question:</p> <p>چرا وجود غده پروستات در سیستم ادراری مردان، از نظر متن، می‌تواند به عنوان یک مزیت نقی شود؟</p> <p>Answer:</p> <p>غده پروستات مانند یک مانع فیزیکی عمل می‌کند که میزراه مردانه را طولانی‌تر و پیچیده‌تر می‌سازد. این پیچیدگی و طولانی‌تر شدن مسیر، ورود عامل بیماری‌زا از خارج بدن را دشوارتر می‌کند.</p>	A
<p>Counterfactual:</p> <p>Question:</p> <p>اگر اسفنکتر خارجی مثانه به درستی کار نمی‌کرد، چه پیامد ناخواسته و مهمی برای فرد به وجود می‌آمد؟</p> <p>Answer:</p> <p>فرد قادر به کنترل ارادی دفع ادرار خود نمی‌بود و مایعات بدن به صورت غیررادی و ناخواسته خالیه می‌شندند. این امر می‌توانست زندگی روزمره و اجتماعی او را با چالش‌های جدی مواجه کند.</p>	A

Gemini:

Our body is a marvel of engineering! Imagine the bladder as a reservoir for excess fluids. To empty this organ, there's a crucial tube called the urethra, which plays the main role in discharging urine. Now imagine two different plumbing systems—one for *Mary* and another for *Reza*. In *Mary*'s body, the urethra is very short, about 2.5 to 5 centimeters long. This close proximity to the external environment unfortunately means that "external problems" or infections can more easily and quickly find their way in compared to the male system. In fact, urinary tract infections (UTIs) occur several times more often in women due to this anatomical difference. In *Reza*'s body, however, the situation is different. His urethra is much longer, measuring about 20 to 25 centimeters. This tube has curves and passes through a protective chamber (the prostate gland) and then through an external structure (the penis) before reaching its end and releasing its contents. And one more important detail: right at the beginning of both urethras, there is a smart control valve called the external urethral sphincter. This valve plays a crucial role—like a gatekeeper, it prevents involuntary leakage of urine and gives us full control over when we urinate. This precise design ensures that the bladder empties at the right time and in a controlled manner.

Rate
A

Gemini:

<p>Multiple Choice Question: If the female urethra were as long and complex as the male urethra, which of the following outcomes would be more likely? A) Bodily fluids would be expelled more quickly. B) Women's bodies would be less exposed to urinary tract infections. C) The prostate gland would also develop in women. D) The need for voluntary control of urination would disappear. Answer: B</p>	A
<p>True/False Judgment 1: The short length of the female urethra is one of the main reasons for the increased likelihood of external pathogens entering the urinary system. Answer: True</p>	A
<p>True/False Judgment 2: According to the text, the prostate gland is a key organ responsible for voluntary control of urination in men. Answer: False</p>	A
<p>Cause and Effect: Cause: The presence of an intelligent control valve known as the external urethral sphincter Effect: The human ability to voluntarily control and prevent unintentional discharge of bodily fluids</p>	A
<p>Cloze: Sentence: The structural difference in fluid discharge systems between men and women—especially in urethral length—unfortunately means that infections can “_____.” Answer: find their way in far more easily</p>	A
<p>Explanation: Question: Why is the presence of the prostate gland in the male urinary system considered an advantage according to the text? Answer: The prostate gland acts as a physical barrier that makes the male urethral path longer and more complex. This added complexity and length make it harder for external pathogens to enter the body.</p>	A
<p>Counterfactual: Question: What would happen if the external urethral sphincter didn't function properly? Answer: The person would lose voluntary control over urination, leading to involuntary and unwanted discharge of bodily fluids. This could seriously affect their daily life and social interactions.</p>	A

Gemini:

پیشنهای روز بین شنبه و دوشنبه است، ایزد مهر نگهبان این روز در فرهنگ ایران است. نام این روز در میان سغدیان و مانوی‌ها از به هم پیوستن دو واژه Mir و jmnw ساخته شده‌است. بخش نخست نام سغدی ایزد پیمان یا مهر است و بخش دوم همان واژه روز در زبان فارسی است؛ بنابراین نام پیشنهای در ایران باستان Mir و برگردان این نام به فارسی، مهر روز است. دومین روز هفته در گامشماری خورشیدی و آخرين و هفتمین روز هفته در گامشماری میلادی است. در کشورهای اروپایی و امریکایی این روز تعطیل رسمی هفتگی است. برای بیشتر مسیحیان، پیشنهای به عنوان روز عبادت و استراحت است و به عنوان روز خداوند و روز رستاخیز مسیح در نظر گرفته شده‌است. در اسرائیل، ایالات متحده، کانادا، چین، ژاپن و فیلیپین و همچنین در آمریکای جنوبی، پیشنهای اولین روز هفته است.

Rate	پیشنهای... روزی خاص که همیشه حس و حال خودش را دارد. دقیقاً بین شنبه شلوغ و دوشنبه پریاهو قرار گرفته، مثل پک پل آرامش. در فرهنگ غنی ما ایرانی‌ها، از دیرباز ایزد مهر، نگهبان این روز بوده، گویی سایه طفش را بر آن گسترانیده است. تصویرش را بکنید، مردمان باستانی مثل سغدی‌ها و مانوی‌ها، این روز را با ترکیب دو کلمه «میر» و «جنو» نامگذاری کرده بودند. «میر» برای آن‌ها ایزد پیمان یا همان مهر بود و «جنو» هم درست مثل «روز» ما در فارسی، به معنای «روز» می‌آمد. به همین خاطر، در ایران باستان نام پیشنهای «میرجنو» بود که ترجمه فارسی آن می‌شود «مهر روز».
A	جالب اینجاست که در گامشماری خورشیدی ما، پیشنهای دومین روز هفته به حساب می‌آید، اما در گامشماری میلادی، آخرين و هفتمین روز هفته است؛ یک تناقض شیرین تقویی؛ در سفر به کشورهای اروپایی و امریکایی بارها دیده‌ام که این روز تعطیل رسمی هفتگی است و خیابان‌ها از امشی خاص دارند. برای بیشتر مسیحیان، پیشنهای روز خداوند و استراحت است، روزی که آن را روز خداوند و روز رستاخیز مسیح می‌دانند و با احترام خاصی به کلیسا می‌روند یا در خانه آرام می‌گیرند. اما ماجرا‌ای پیشنهای به اینجا ختم نمی‌شود؛ اگر در اسرائیل، ایالات متحده، کانادا، چین، ژاپن، فیلیپین یا حتی آمریکای جنوبی زندگی کنید، متوجه می‌شود که پیشنهای اولین روز هفته است و آغازگر همه چیز! گویی هر کجا که باشی، پیشنهای با داستان‌ها و نقش‌های متفاوت‌شون، همیشه حضوری پررنگ دارد.

Gemini:

Multiple Choice Question: چرا یکشنبه در متن به عنوان «پل آرامش» توصیف شده است؟ A) چون ایزد مهر نگهبان این روز است. B) زیرا بین دو روز پرمشغله شنبه و دوشنبه قرار گرفته و خود روزی برای استراحت است. C) به خاطر قدمت تاریخی آن در فرهنگ باستانی. D) چون در گاهشماری میلادی آخرین روز هفته است. Answer: B	A
True/False Judgment 1: آرامش خیابان‌ها در کشورهای اروپایی و آمریکایی در روز یکشنبه، شبانه‌های از کار نکردن عمومی و روحیه آرام این روز است. Answer: (درست) (True)	A
True/False Judgment 2: در متن آمده که یکشنبه در همه کشورها یک روز کاری و پرمشغله است. Answer: (نادرست) (False)	A
Cause and Effect: Cause: قرار گرفتن یکشنبه بین دو روز شلوغ و کاری Effect: ایجاد حس آرامش و فرصت استراحت در میان روزهای پرنش هفته	A
Cloze: Sentence: در بسیاری از فرهنگ‌ها، یکشنبه به عنوان «_____» شناخته می‌شود که فرصت مناسبی برای استراحت ذهن و بدن فراهم می‌آورد. Answer: پل آرامش	A
Explanation: Question: چرا موقعیت زمانی یکشنبه در تقویم هفتگی باعث شده تا از آن به عنوان روزی برای استراحت یاد شود؟ Answer: چون یکشنبه بین شنبه (شروع پرشور هفته) و دوشنبه (روز بازگشت جدی به کار) قرار دارد و از نظر روانی و زمانی فرصنی برای توقف، استراحت و تجدید انرژی فراهم می‌آورد.	A
Counterfactual: Question: اگر یکشنبه نیز مانند دیگر روزهای هفته کاری و شلوغ بود، چه تأثیری بر زندگی و سلامت روان افراد می‌گذاشت؟ Answer: افراد فرصت کمتری برای استراحت و تجدید قوا داشتند، که ممکن بود منجر به افزایش خستگی مزمن، کاهش بهره‌وری و فشار روانی بیشتر شود.	A

Gemini:

<p>C (in English: C) is a general-purpose, compiled, mid-level, structured, imperative, and procedural programming language, developed in 1972 by Dennis Ritchie at Bell Labs. Sunday is the day between Saturday and Monday, and in Iranian culture, Mithra (the deity of kindness) is the guardian of this day. The name of this day among the Sogdians and Manicheans is formed by combining two words: <i>Mir</i> and <i>jmnw</i>. The first part, <i>Mir</i>, refers to the deity of covenant or Mithra, and the second part, <i>jmnw</i>, means “day” in Persian. Therefore, the name Sunday in ancient Iran was <i>Mir jmnw</i>, which translates to “Mithra’s day” in Persian. Sunday is the second day of the week in the solar calendar and the last, seventh day of the week in the Gregorian calendar. In European and American countries, this day is a weekly public holiday. For most Christians, Sunday is considered a day of worship and rest, regarded as the Lord’s day and the day of Christ’s resurrection. In Israel, the United States, Canada, China, Japan, the Philippines, and also in South America, Sunday is the first day of the week.</p>	<p>Rate</p>
<p>Sunday... a special day that always carries its own unique feeling. It sits exactly between the busy Saturday and the hectic Monday, like a bridge of calm. In our rich Iranian culture, from ancient times, Mithra—the deity of kindness—has been the guardian of this day, as if spreading a protective shadow over it. Imagine this: ancient peoples like the Sogdians and Manicheans named this day by combining two words—“Mir” and “Jammū.” For them, “Mir” meant the deity of covenant or Mithra, and “Jammū,” just like our Persian word for “day,” meant “day.” That’s why in ancient Iran, Sunday was called “Mirjamnū,” which translates to “Mithra’s day” in Persian. Interestingly, in our solar calendar, Sunday is considered the second day of the week, but in the Gregorian calendar, it is the last and seventh day of the week—a charming calendar contradiction! During my travels in European and American countries, I have often noticed that Sunday is an official weekly holiday, and the streets hold a special kind of calm. For most Christians, Sunday is a day of worship and rest, regarded as the Lord’s day and the day of Christ’s resurrection, when people go to church with great respect or rest quietly at home. But the story of Sunday doesn’t end there; if you live in Israel, the United States, Canada, China, Japan, the Philippines, or even South America, you will notice that Sunday is the first day of the week—the beginning of everything! It seems wherever you are, Sunday—with its stories and different roles—always has a prominent presence.</p>	<p>A</p>

Gemini:

<p>Multiple Choice Question: Why is Sunday described as the "Bridge of Peace" in the text? A) Because Mithra (the deity of kindness) is the guardian of this day. B) Because it is placed between two busy days, Saturday and Monday, and is itself a day for rest. C) Due to its historical significance in ancient culture. D) Because in the Gregorian calendar it is the last day of the week. Answer: B</p>	<p>A</p>
<p>True/False Judgment 1: The calmness of the streets in European and American countries on Sunday is a sign of general non-working and the peaceful spirit of this day. Answer: True</p>	<p>A</p>
<p>True/False Judgment 2: According to the text, Sunday is a busy and working day in all countries. Answer: False</p>	<p>A</p>
<p>Cause and Effect: Cause: Sunday is positioned between two busy and working days. Effect: It creates a sense of calm and provides an opportunity for rest amid the hectic days of the week.</p>	<p>A</p>
<p>Cloze: Sentence: In many cultures, Sunday is known as the "____" which offers a suitable opportunity for mental and physical rest. Answer: Bridge of Peace</p>	<p>A</p>
<p>Explanation: Question: Why does the timing of Sunday in the weekly calendar cause it to be considered a day of rest? Answer: Because Sunday falls between Saturday (the energetic start of the week) and Monday (the serious return to work), providing a psychological and temporal opportunity to pause, rest, and renew energy.</p>	<p>A</p>
<p>Explanation: Question: Why does the timing of Sunday in the weekly calendar cause it to be considered a day of rest? Answer: Because Sunday falls between Saturday (the energetic start of the week) and Monday (the serious return to work), providing a psychological and temporal opportunity to pause, rest, and renew energy.</p>	<p>A</p>

G Experimental Setup

G.1 Training Configuration

We fine-tune six instruction-tuned base models (MISTRAL-SMALL-24B, MISTRAL-7B, MISTRAL-NEMO, and LLAMA-3 variants with 1–8 B parameters) on the CS_Farsi corpus (100% of the train split). All experiments are run with DEEPSPEED ZeRO-3 and Flash-Attention 2 on three NVIDIA RTX A6000 GPUs (48 GB each); bfloat16 mixed precision is enabled.

Effective batch: 4 sequences/GPU \times 2 grad-
accum. steps \Rightarrow 8 sequences (\approx 16k tokens
with packing).

Max sequence length: 2,048 tokens (samples
are packed).

Epochs: 3.

Optimizer: AdamW with cosine decay; initial
LR 1×10^{-4} ; weight decay 1×10^{-4} ; no
warm-up.

LoRA: rank $r = 128$, scaling $\alpha = 256$,
dropout 0.10.

Adapted modules: q_proj, k_proj, v_proj,
o_proj, down_proj, up_proj, gate_proj,
embed_tokens, lm_head.

Regularisation: gradient clipping at 1.0.

G.2 Inference Configuration

Decoding uses nucleus sampling with tempera-
ture 0.8, top_p = 0.95, and top_k = 50; a repetition
penalty of 1.1 mitigates degeneration.

Table 11: Wall-clock training time per base model ($3 \times$ A6000).

Base model	Parameters	Runtime
MISTRAL-SMALL-24B-INSTRUCT-2501	24 B	13 h 22 m
MISTRAL-7B-INSTRUCT-v0.3	7 B	9 h 37 m
MISTRAL-NEMO-INSTRUCT-2407	12 B	5 h 48 m
LLAMA-3.1-8B-INSTRUCT	8 B	4 h 03 m
LLAMA-3.2-3B-INSTRUCT	3 B	2 h 01 m
LLAMA-3.2-1B-INSTRUCT	1 B	55 m