PERCOR: Evaluating Commonsense Reasoning in Persian via Multiple-Choice Sentence Completion

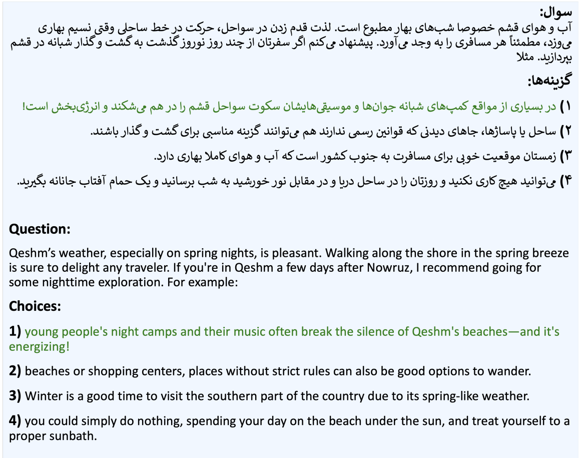
Morteza Alikhani[†] **Mohammadtaha Bagherifard**[†*] **Erfan Zinvandi**
Mehran Sarmadi
MCINEXT
`taha.bagheri98@gmail.com`
`{morteza.alikhani95, e.zeynvandi1376, mehran.sarmadi99}@sharif.edu`

Abstract

We introduced **PERCOR**—***Per****sian* ***Co****mmonsense* ***R****easoning*—the first large-scale Persian benchmark for commonsense reasoning. PERCOR contains 106K multiple-choice sentence-completion problems drawn from more than forty news, cultural and other web sources. We adopt a linguistically grounded, conjunction-based segmentation strategy to generate coherent prefix–continuation pairs. To create challenging distractors, we propose **DRESS-AF**—***D****istractor* ***R****anking via* ***E****mbedding* ***S****imilarity* ***S****coring and* ***A****dversarial* ***F****iltering*—a generation-free adversarial filtering method that selects distractors from the pool of gold continuations while maximising model confusion. Human annotators score 89% on PERCOR, while `OpenAI-o3` achieves the highest performance at 92.18%, followed closely by `Claude-Sonnet-3.7` (91.17%). The strongest open-source model, `DeepSeek-R1`, reaches 82.51%, underscoring both the dataset's difficulty and the remaining performance gap in Persian commonsense reasoning. We further show that DRESS-AF transfers to the English HellaSwag benchmark, increasing its difficulty without hurting human solvability. The dataset is available at `https://huggingface.co/datasets/MCINext/PerCoR`.

1 Introduction

Commonsense reasoning is a critical capability in natural language understanding, enabling models to draw inferences, disambiguate meaning, and interpret implicit knowledge. While large language models (LLMs) have shown remarkable progress across various tasks, their performance on commonsense reasoning—particularly in structured formats like multiple-choice sentence completion—remains limited (Sap et al., 2020b).



سوال:
آب و هوای قشم خصوصا شب‌های بهار مطبوع است. لذت قدم زدن در سواحل، حرکت در خط ساحلی وقتی نسیم بهاری
می‌وزد، مطمئناً هر مسافری را به وجد می‌آورد. پیشنهاد می‌کنم اگر سفرتان از چند روز نوروز گذشت به گشت و گذار شبانه در قشم
بپردازید. مثلا

گزینه‌ها:

۱) در بسیاری از مواقع کمپ‌های شبانه جوان‌ها و موسیقی‌هایشان سکوت سواحل قشم را در هم می‌شکند و انرژی‌بخش است!

۲) ساحل یا پاساژها، جاهای دیدنی که قوانین رسمی ندارند هم می‌توانند گزینه مناسبی برای گشت و گذار باشند.

۳) زمستان موقعیت خوبی برای مسافرت به جنوب کشور است که آب و هوای کاملا بهاری دارد.

۴) می‌توانید هیچ کاری نکنید و روزتان را در ساحل دریا و در مقابل نور خورشید به شب برسانید و یک حمام آفتاب جانانه بگیرید.

Question:

Qeshm's weather, especially on spring nights, is pleasant. Walking along the shore in the spring breeze is sure to delight any traveler. If you're in Qeshm a few days after Nowruz, I recommend going for some nighttime exploration. For example:

Choices:

1) young people's night camps and their music often break the silence of Qeshm's beaches—and it's energizing!

2) beaches or shopping centers, places without strict rules can also be good options to wander.

3) Winter is a good time to visit the southern part of the country due to its spring-like weather.

4) you could simply do nothing, spending your day on the beach under the sun, and treat yourself to a proper sunbath.

Figure 1: An example from the PERCOR dataset. The passage discusses the pleasant spring weather in Qeshm and recommends nighttime exploration. The correct answer (written in Green) refers to night camps and music breaking the beach's silence, while other options, though plausible in isolation, lack relevance to the immediate context.

To benchmark and improve this ability, several datasets such as SWAG (Zellers et al., 2018), HellaSwag (Zellers et al., 2019), and CommonsenseQA (Talmor et al., 2019) have been proposed. However, these benchmarks are overwhelmingly English-centric, leaving a significant gap in resources for evaluating and improving commonsense reasoning in low-resource languages.

Despite recent progress in Persian NLP through resources such as PARSINLU (Khashabi et al., 2021), PersianQA (Ayoubi, 2021), and PQUAD (Darvishi et al., 2023), Persian remains a low-resource language for high-level reasoning tasks, particularly commonsense inference. This leaves a significant gap in evaluating and advancing structured reasoning capabilities in this language. To address this limitation, we introduce **PERCOR**—the first large-scale Persian commonsense reasoning dataset in multiple-choice sentence–completion for-

[†]Equal contribution.
[*]Corresponding author.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2205–2224
December 20-24, 2025 ©2025 Association for Computational Linguistics

mat. Constructed from over 40 diverse Persian websites, PERCOR captures a broad range of domains and linguistic styles. We formulate each instance as a sentence prefix followed by four completion candidates: one correct and three distractors. Instead of relying on simple rule-based methods for sentence segmentation, we generate sentence–completion pairs by splitting at conjunctions, promoting natural flow and semantic coherence. This connective-driven formulation is related to prior work that uses discourse markers for inference and sentence completion (Bhargava and Ng, 2022), but PERCOR applies this strategy at scale to Persian and across diverse, non-narrative domains. Unlike oversimplified strategies such as the one employed in SWAG (Zellers et al., 2018), which relies on temporally grounded data like video captions, our conjunction-based approach is applicable to a wide range of textual sources. This enables broader domain coverage and greater variability in sample length, enhancing both the diversity and richness of the dataset.

We further propose a novel distractor selection strategy, DRESS-AF, which is a combination of Adversarial Filtering (AF) (Zellers et al., 2019) and Embedding-based ranking (Liang et al., 2018; Chiang et al., 2022) methods. DRESS-AF avoids LLM-based generations—thus sidestepping associated biases—and instead ranks completions using embedding-based similarity metrics. These scores are adversarially tuned to maximise model confusion using Bayesian optimisation over a development set, yielding difficult yet human-solvable distractors.

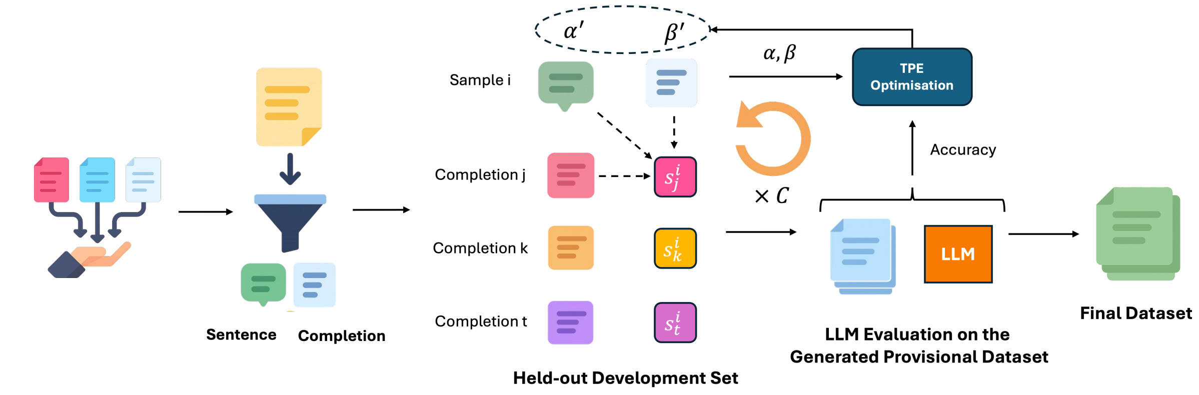
PERCOR specifically targets discourse-level and contextual commonsense reasoning. To select the correct continuation, a model must infer the semantic relation introduced by the conjunction (e.g., causal, contrastive, elaborative) and identify which candidate maintains coherence with the preceding context. Distractors are intentionally plausible in isolation but violate these discourse constraints, requiring models to integrate implicit world knowledge with local coherence cues rather than relying on shallow lexical matching.

An example is shown in Figure 1 illustrating a key aspect of PERCOR dataset—candidates are intentionally context-sensitive. While all options may appear semantically valid in isolation, only one logically follows from the passage. In this case, the mention of "nighttime exploration" cues the correct choice, requiring the model to interpret implicit temporal references to succeed.

In summary, our key contributions are as follows: **(1)** we introduce **PERCOR**, the first large-scale Persian commonsense reasoning dataset in a multiple-choice sentence–completion format, spanning diverse domains and linguistic styles; **(2)** we propose a conjunction-based extraction method that enables natural and semantically coherent sample generation from non-temporal texts; **(3)** we present **DRESS-AF**, a language-agnostic, embedding-based distractor generation approach that incorporates adversarial filtering to produce challenging yet human-solvable distractors—without relying on generative models; and **(4)** we benchmark a broad set of state-of-the-art open- and closed-source LLMs on PerCoR, establishing strong empirical baselines for future work.

2 Related Work

Commonsense Reasoning Datasets. Numerous English benchmarks have been introduced to evaluate commonsense reasoning in multiple formats. SWAG (Zellers et al., 2018) and HELLASWAG (Zellers et al., 2019) pose multiple-choice sentence completion tasks based on narrative or descriptive contexts. HellaSwag, in particular, uses adversarial filtering to create distractors that are challenging for language models but easily solvable by humans. Cloze-style narrative continuation tasks such as the STORY CLOZE TEST (Mostafazadeh et al., 2016) also involve selecting a plausible continuation, though they focus on narrative coherence rather than connective-driven discourse inference. Other benchmarks such as WINOGRANDE (Sakaguchi et al., 2021), COMMONSENSEQA (Talmor et al., 2019), OPENBOOKQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), COSMOS (Huang et al., 2019), and SOCIAL IQA (Sap et al., 2019) cover a variety of commonsense dimensions, including physical reasoning, social dynamics, and multi-hop inference. More recent efforts include GLUCOSE (Mostafazadeh et al., 2020), a dataset of causal explanations in short narratives, annotated across ten dimensions of inferential knowledge; COM2SENSE (Singh et al., 2021), which evaluates a model's ability to discriminate between true and false commonsense statements in complementary pairs; and COMMONSENSEQA 2.0 (Talmor et al., 2021), an adversarially curated yes/no question dataset designed to be difficult for large language models while remaining easy for humans.



(a) Data Collection (b) Sentence-Completion Creation (c) Distractor Generation

Figure 2: Overview of our dataset construction and distractor generation pipeline. The process consists of: (a) collecting diverse Persian text data, (b) creating and filtering sentence-completion pairs, and (c) generating challenging multiple-choice distractors using **DRESS-AF**.

Despite substantial progress, these benchmarks are primarily designed for English, leaving a gap in resources for other languages.

Distractor Generation Techniques. Creating high-quality distractor candidates is crucial for constructing reliable multiple-choice datasets (Alhazmi et al., 2024). *Adversarial filtering* (AF), used in SWAG, HellaSwag, and WinoGrande, iteratively removes easy distractors using a discriminator model, resulting in semantically challenging options. Alternatively, *retrieval-based methods* select distractors from external corpora or knowledge graphs, ensuring topical relevance (Ren and Zhu, 2021b). Recent work extends this by incorporating topic models to filter noisy candidates from knowledge graphs like Probase (Ren and Zhu, 2021a). *Embedding-based ranking* selects distractors based on similarity in embedding space (Liang et al., 2018; Chiang et al., 2022), while *retrieval-augmented generation* leverages retrieved passages and knowledge triplets to guide large language models in producing diverse distractors (Chen et al., 2023). Our proposed method, DRESS-AF, combines these principles: we rank gold completions using embedding-based similarity scoring and adversarially optimise parameters to select distractors that maximise model confusion.

Persian NLP Resources. Recent years have seen a growing body of work on Persian NLP, but most resources target core tasks such as machine translation, sentiment analysis, and reading comprehension. ParsiNLU (Khashabi et al., 2021) includes benchmarks for Persian NLI, QA, and sentiment classification. FarsTail (Amirkhani et al., 2023) is a natural language inference dataset, while PQuAD (Darvishi et al., 2023) provides large-scale reading comprehension benchmarks in the SQuAD format. For more open-ended reasoning, PerCQA (Jamali et al., 2022) is a community QA dataset compiled from Persian web forums, consisting of 989 real-world questions and over 21k answers, designed for tasks like answer selection and ranking. Although these resources enable evaluation of Persian understanding and reasoning, they do not address commonsense reasoning specifically. To the best of our knowledge, our work presents the first large-scale Persian commonsense reasoning dataset, addressing a significant gap in low-resource language evaluation.

3 The PerCoR Dataset

We adopt a three-stage pipeline to create the PerCoR dataset: (1) **Data Collection**, in which raw text segments are gathered from diverse sources; (2) **Sentence–Completion Creation**, where sentence–completion pairs are generated using our novel conjunction-based method; and (3) **Distractor Generation**, where we apply our proposed DRESS-AF algorithm to select challenging distractor candidates for each instance. An overview of this pipeline is illustrated in Figure 2.

3.1 Data Collection

To construct our dataset, we begin by collecting a diverse set of paragraphs spanning a broad range of topics, ensuring that meaningful sentence–completion pairs can be extracted for multiple-choice commonsense evaluation. For this purpose, we leverage the Corpesia corpus (Sarmadi et al., 2025), a large-scale resource built by crawling the main content (excluding advertisements and irrelevant sections) from a wide variety of Persian websites. The raw data in Corpersia is cleaned through rule-based filtering to remove boilerplate artifacts—such as author names, timestamps, and footers—while preserving the original paragraph structure and maintaining document-level segmentation, including title identification.

To generate sentence–completion pairs, we select a subset of websites from Corpersia that cover a broad spectrum of topics (detailed in Section 4.1). We discard any paragraph with fewer than 50 characters to ensure the textual quality and context richness. Finally, we sample up to 200,000 paragraphs from each selected website to be used as the source for extracting sentence–completion pairs.

3.2 Sentence–Completion Creation

Rather than relying on conventional techniques such as those employed in SWAG (Zellers et al., 2018), which depend on temporally coherent data (e.g., video captions), we adopt a linguistically grounded strategy based on conjunctions to extract sentence–completion pairs from static text. Specifically, we begin by curating a list of 49 high-frequency conjunctions in Persian. To ensure consistency and reduce sparsity, we remove conjunctions that appear fewer than 500 times in the corpus. Importantly, all excluded items have semantically equivalent counterparts in the retained set, preserving the expressivity of the conjunction space. The final list of conjunctions, along with their English translations, is presented in Figure 8. To maintain balanced representation across conjunctions and avoid dominance by high-frequency items, we sample up to 4,000 instances per connective. If a conjunction occurs fewer than 4,000 times, we include every instance. For semantically ambiguous conjunctions—those that may not always function as true connectives in contexts—we increase our oversampling multiplier so that the filtered data retains a sufficient number of valid usages.

To ensure that the sentence–completion split occurs at an informative and coherent boundary, we define a valid character span within which conjunctions are considered—ranging from a minimum of 50 to a maximum of 250 characters from the start of the paragraph. The lower bound ensures that the prefix contains sufficient context for prediction, while the upper bound prevents overly long or semantically overloaded prefixes. Once a valid conjunction is found within this range, we check the character length of the clause following it. If the length is below a threshold of 150 characters, the paragraph is split at that conjunction to form a sentence–completion pair. Otherwise, the search continues with other conjunctions in the span. This ensures that the completion segment remains concise and focused.

To further validate the quality of extracted pairs , we perform a lightweight filtering step using the `GPT-4o-mini` model. Specifically, the model is used for two binary classification checks: (1) verifying that the identified conjunction functions as a true discourse connective (since some Persian conjunctions may be contextually ambiguous), and (2) ensuring that the completion segment is a syntactically and semantically complete sentence. Since the model is only used for verification, not generation, it does not introduce generation-related biases into the data. Additional details regarding this filtering process are provided in Appendix A.1.

3.3 Distractor Generation

To avoid introducing any biases associated with language model generations, we select distractor options from the set of gold completions belonging to other samples, rather than generating them via an LLM. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be the embedding of the sentence and completion, respectively. We define a score s^i_j, representing the suitability of completion$_j$ as a candidate option for sentence$_i$, as follows:

$$\begin{aligned} s^i_j &= \alpha \cos(\mathbf{x}_i, \mathbf{y}_j) + \beta \cos(\mathbf{y}_i, \mathbf{y}_j) \\ &\quad + (1 - \alpha - \beta) \cos(\mathbf{z}_i, \mathbf{y}_j), \end{aligned} \tag{1}$$

where $\alpha, \beta \in [0, 1]$ are tunable coefficients that balance the contributions of each similarity term, $\cos(\cdot, \cdot)$ denotes cosine similarity, and $\mathbf{z}_i$ refers to the embedding of concatenation of the sentence and its gold completion. Using a held-out development set, we compute s^i_j for each sample pair i within the development set and all candidates j within the whole data (not only in the development set). Based on these scores, we sort the candidates in descending order, exclude the gold completion $\mathbf{y}_i$, and uniformly sample three distractors from the next k-best candidates. This process yields a 4-way multiple-choice instance for each sentence in the held-out set, constructed dynamically according to the current values of α and β.

We optimize α and β via adversarial filtering: for a given (α, β), we build the provisional dataset from the held-out development set, measure the accuracy of an LLM on it, and use that accuracy as the objective in a Tree-structured Parzen Estimator (TPE) Bayesian optimization over c trials. Although the search space is low-dimensional, TPE is known for its strong empirical performance and sample efficiency in hyperparameter tuning tasks, and has been widely adopted in AutoML and deep learning optimisation pipelines (Watanabe, 2023).[1]. The optimal (α^*, β^*) are then used to generate our final dataset. To construct the PERCOR dataset, we set $c = 30$ and $k = 20$. We employed the HAKIM embedding model (Sarmadi et al., 2025), as it demonstrated the best performance on FAMTEB (Zinvandi et al., 2025), a comprehensive benchmark for Persian text embeddings.

We refer to this method as **DRESS-AF** (*Distractor Ranking via Embedding Similarity Scoring and Adversarial Filtering*). DRESS-AF constructs multiple-choice questions by scoring all candidate completions using the embedding-based metric defined above, and then adversarially optimising the scoring parameters to select the most challenging distractors. Importantly, the adversarial nature of DRESS-AF ensures that the selected distractors increase question difficulty—but it does not guarantee overall dataset quality or standardness. In practice, two hyperparameters play a key role in adjusting the difficulty: c, the number of optimisation trials, and k, the number of top-ranked distractor candidates (after excluding the gold completion) from which three distractors are randomly sampled. While DRESS-AF aims to generate difficult examples for evaluation, human oversight may still be required to discard samples that are excessively ambiguous or unsolvable, ensuring that the final dataset remains reliable and informative.

We hypothesise that the set of gold completions across all samples is sufficiently diverse to serve as a reliable pool of distractor candidates. This assumption enables us to avoid synthetic generation altogether and sidestep potential biases introduced by LLM outputs. In Section 4, we empirically validate this hypothesis by showing that several strong LLMs consistently achieve below 80% accuracy on our dataset. This confirms the overall challenge posed by the distractors selected via DRESS-AF. Furthermore, evaluations conducted by human annotators on a subset of the data yield accuracies around 90%, providing additional evidence that the questions are both plausible and solvable, albeit non-trivial.

4 Experiments

We structure our experiments in three phases: first, we analyse the PERCOR dataset by examining token-length distributions and covered topics and domain; second, we evaluate DRESS-AF's ability to craft challenging distractors for the sentence-completion pairs in the HellaSwag dataset, demonstrating the generality of our method in generating strong distractors without relying on generative models, and also its applicability beyond Persian; third, we benchmark 32 large language models on the dataset in a zero-shot setting to gauge their out-of-the-box performance. Further experiments regarding the effect of input length and also few-shot evaluation are provided in Appendix B.2, B.3.

4.1 Dataset Statistics

The dataset is divided into three splits: training (86,217 samples), validation (10,000 samples), and test (10,000 samples). Each sample consists of an uncompleted text and four candidate completions. The average sentence length is 129.23 characters and 41.78 tokens, while the average completion length is 93.24 characters and 30.08 tokens. Completion statistics are computed by first averaging the length (in characters and tokens) across the four candidates within each sample, and then taking the mean over all samples. Token lengths are calculated using the *GPT-4o-mini* tokeniser via the `tiktoken` library (OpenAI, 2023).

To ensure linguistic and topical diversity in our dataset, we collected raw Persian text data from over 40 distinct websites spanning a broad range of domains. These include news and current affairs (e.g., ISNA, KhabarOnline, YJC), technology and digital media (e.g., Digiato, Zoomit), religion and culture (e.g., Hawzah, WikiShia, Wiki Ahlolbait), lifestyle and health (e.g., NiniSite, Doctoreto, Namnak), economy and business (e.g., EqtesadOnline, Ecoiran, Digikala Mag), travel and leisure (e.g., Hamgardi, Alibaba), education and self-improvement (e.g., Fidibo, Taaghche, Motamem), and sports and entertainment (e.g., Varzesh3, VIPofilm). In addition, user-generated content platforms like Virgool contribute informal

[1] We employed the implementation of TPE in the `Optuna` library (Akiba et al., 2019).

and diverse writing styles. This domain variety enables broad coverage of content structures, writing registers, and topics, making the dataset a representative resource for real-world commonsense reasoning in Persian.

4.2 Effectiveness of DRESS-AF in Distractor Generation

4.2.1 PerCoR Dataset

To evaluate the effectiveness of our proposed distractor generation method during the construction of the dataset, we track the performance of the `GPT-4o-mini` model on the provisional datasets constructed during the optimization process. Specifically, we run $c = 30$ trials, where in each trial we use a different pair of (α, β) coefficients to generate distractors based on the scoring function defined in Section 3. Our goal is to adversarially reduce the model's accuracy—i.e., to identify distractor settings that make the multiple-choice task more challenging. Among the c trials, we select the (α^*, β^*) pair corresponding to the lowest model accuracy for use in the final dataset construction.

Figure 3 shows the model's accuracy across the 30 trials. For the first 10 trials, we use random initialization to encourage exploration; from trial 11 onward, we apply the Tree-structured Parzen Estimator (TPE) algorithm for guided search. We plot the accuracy along with a rolling standard deviation (window size = 3) to visualize exploration dynamics. As seen, the variance is initially high due to random sampling, then decreases as the optimization converges. The lowest observed accuracy occurs at trial 20, indicating the most adversarial configuration found by DRESS-AF.

4.2.2 HellaSwag Dataset

To further demonstrate the effectiveness and generality of DRESS-AF in generating challenging distractor candidates without introducing generation-induced biases from LLMs, we apply the method to a non-Persian benchmark: the HellaSwag dataset (Zellers et al., 2019). Specifically, we take the validation split of the HellaSwag dataset, then use its sentence–completion pairs (i.e., the context and gold ending) as inputs to DRESS-AF, showcasing the method's language-agnostic applicability.

To evaluate the extent to which DRESS-AF allows control over distractor difficulty, we construct two new variants of HellaSwag. In the first (harder) version, for each sample, we randomly sample three distractors from the top 10 highest-scoring candidates based on the embedding similarity score (excluding the gold completion). In the second (easier) version, we exclude the top 3 candidates (and the gold completion if it is not among them), then sample three distractors from the next top 20. For both versions, we run $c = 30$ optimisation trials to find the best (α, β) parameters via the DRESS-AF procedure. For both variations, we employed Jina-v3 (Sturua et al., 2024) as the embedding model.

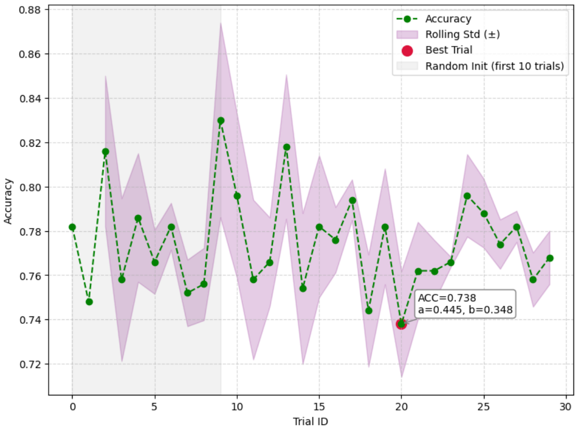


Figure 3: Accuracy of GPT-4o-mini on the provisional dataset, during the construction of the PerCoR dataset. DRESS-AF tries to find the best coefficients within 30 trials. The first 10 trials use random sampling, followed by TPE-based search. The lowest accuracy (trial 20) corresponds to the selected distractor configuration.

Figure 4 shows the accuracy of `GPT-4o-mini` on provisional datasets over the 30 trials during the tuning process of (α, β). The observed trend resembles the Persian setup in Figure 3: during the initial 10–15 randomly sampled trials, variance is high due to exploration; afterward, performance stabilises as TPE converges. Using the best-found (α, β), we finalise the two dataset variants. We then evaluate both closed-source (`GPT-4o-mini`) and open-source (`Gemma-3-27B-it`) models on these variants, as well as on the original HellaSwag, to assess how distractor difficulty affects performance.

Figure 5 presents model performance across three versions of the HellaSwag dataset: the original, an easier variant, and a harder one—both constructed using DRESS-AF. As expected, accuracy decreases as the distractor difficulty increases, demonstrating the method's effectiveness in producing more challenging distractors. Notably, both `GPT-4o-mini` and `Gemma3-27B-it` exhibit the lowest accuracy on the harder variant, indicating that DRESS-AF successfully identifies distractors that

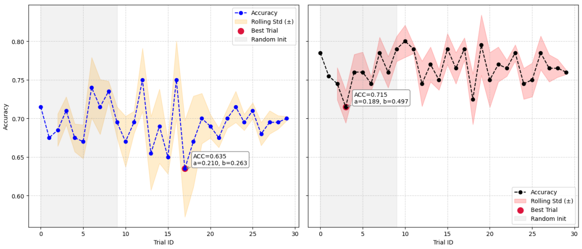


Figure 4: Accuracy of GPT-4o-mini on the provisional dataset across 30 trials during DRESS-AF optimisation on sentence–completion pairs from HellaSwag. The left plot corresponds to the harder version, with distractors sampled from the top 10 candidates. The right plot corresponds to the easier version, where the top 3 candidates are excluded and distractors are sampled from the next top 20.

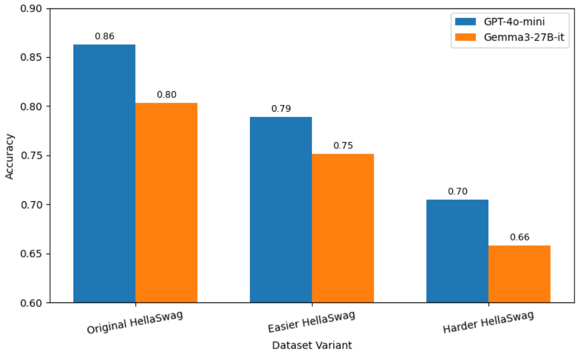


Figure 5: Accuracy of `GPT-4o-mini` and `Gemma3-27B-it`, representing closed- and open-source models respectively, across three HellaSwag variants. DRESS-AF was used to generate distractors for the easier and harder variants.

are more confounding for models.

The performance difference between the easier and harder versions can be attributed to the distractor sampling strategy. In the easier variant, we exclude the top three most confounding candidates and then randomly select from the next top 20. This design favours broader semantic differences between gold and distractor completions. In contrast, for the harder variant, we randomly sample distractors from the top 10 candidates, making the distractors more semantically similar to the correct answer and hence more difficult.

While increasing distractor difficulty is desirable, it is equally important that the dataset remain answerable and free of generation-induced artifacts. To assess this, we conduct a human evaluation on a 200-sample subset of each variant. On the original HellaSwag, human accuracy is 90%. This remains nearly unchanged on the easier variant (89.5%) and decreases only modestly on the harder one (83%). In contrast, model accuracy drops much more sharply: for `Gemma-3-27B-it` and `GPT-4o-mini`, performance decreases by 5–7 points on the easier set and by 14–16 points on the harder set (Figure 5). This asymmetry indicates that models were partially exploiting stylistic or fluency artifacts present in the original LM-generated distractors—artifacts that humans did not rely on. By selecting distractors from real human-written continuations, DRESS-AF reduces such generation-related biases and forces models to rely more on contextual and commonsense reasoning. At the same time, the modest change in human accuracy confirms that the resulting distractors remain human-solvable. In sum, DRESS-AF increases difficulty while maintaining dataset integrity and provides a more robust evaluation signal than generation-based AF methods.

4.3 Model Results on PerCoR

To assess the *out-of-the-box* commonsense abilities of modern LLMs in Persian, we evaluated 12 closed-source and 20 open-source models in a **zero-shot, multiple-choice** setting. Each model was prompted in Persian to return *only* the index of the correct option. We report: (1) **Strict Accuracy**—exact match on the raw output, and (2) **Post-Processed Accuracy**—after applying a simple regex to extract the final digit `1–4`, recovering correct answers when extra justification is included. The model results appear in Table 1.

Overall, closed-source models dominate: `OpenAI-o3` (OpenAI, 2025b) tops the leaderboard at **92.18 %**, followed by `Claude-3.7-Sonnet` (Anthropic, 2025) (91.17 %) and `GPT-4.1` (OpenAI, 2025a) (88.39 %); the best open-source checkpoint, `DeepSeek-R1` (DeepSeek-AI et al., 2025a), reaches **82.51 %**, trimming the gap to roughly 10%, while most open-source peers fall between 60 % and 80 %. Human majority-vote accuracy on PerCoR is 89 % (details in Appendix B.4), so only `o3` and `Sonnet` currently exceed non-expert annotators performance. Despite strong aggregate performance, top-performing models still exhibit occasional failures on nuanced reasoning cases—several examples are provided in Appendix **??**.

Formatting sensitivity is revealed by the gaps between Strict and Post-Processed accuracy: e.g., `GPT-4o` (OpenAI, 2024) from 78.32% to 86.65% (+8.3%), `LLaMA-3.3-70B` (Grattafiori et al., 2024) from 11.23% to 79.56% (+68.3%), `Aya-Expanse-32B` (Dang et al., 2024) from 5.85% to 63.27% (+57.4%), and `DeepSeek-V3` (DeepSeek-AI et al., 2025b) from 51.15% to 82.41% (+31.3%). Large difference indicates that the model often embeds the correct answer in extra prose; shallow post-processing recovers more than 60% of hidden accuracy for some models. In contrast, other models show consistent and similar accuracies, indicating strong adherence to the required output format.

Within individual open-source families, accuracy generally scales with parameter count: the `Gemma3` (Team et al., 2025) series improves from 26% (1B) to 76% (27B), `Qwen-3` (Yang et al., 2025) from 50% (4B) to 76.5% (32B), while `Mistral` (Jiang et al., 2023; Mistral AI, 2025) lags (7B instruct: 30%; 24B "Small-3.1": 69%). `Command A` (Cohere et al., 2025) outperforms its predecessor `Command R` (Cohere, 2024) (79.8 % vs. 60.0 %), likely due to its significantly larger parameter count and improved multilingual alignment—especially in Persian. The `LLaMA-3.2` instruction variants (1B/3B) underperform (<25%), yet the 70B variant, after post-processing, rivals `Gemma3-27B`. These trends confirm that parameter count alone is insufficient; alignment strategy and prompt-format robustness are equally critical on PerCoR.

Closed-source diversity also emerges: `o3 > GPT-4.1 > GPT-4o` suggests benefits from more advanced architecture and reasoning abilities. While `OpenAI-o4-mini` belongs to the same "o-series" family, it underperforms `o3` by a notable margin (85.5 % vs. 92.2 %), potentially due to architectural simplifications or instruction tuning compromises aimed at latency and efficiency. The superior performance of `Gemini-Flash-2.5` over `Flash-2.0` and `Flash-Lite-2.0` (Comanici et al., 2025) reflects incremental training improvements; and `Claude-3.7-Sonnet` (91.2%) outperforming `Claude-3.5-Haiku` (Anthropic, 2024) (71.6%) aligns with Anthropic's published capability tiers.

To further investigate the potential of instruction-tuned open models, we fine-tuned `LLaMA3.3-70B-Instruct` and `Qwen3-32B-Instruct` by applying LoRA (Hu et al., 2022) on the attention layers, leveraging only 10 % of the training data (8,000 samples) for a sequence classification objective. Despite its poor zero-shot performance on strict accuracy (11.23 %), the fine-tuned `LLaMA3.3-70B-Instruct` achieved an accuracy of 86.82 %, while `Qwen3-32B` reached 85.64 %—both surpassing `DeepSeek-R1` (82.51 %) and `DeepSeek-V3` (82.41 %), the strongest open-source models in our zero-shot evaluation. This result highlights the latent capability of instruction-tuned LLMs and demonstrates that even lightweight, resource-efficient fine-tuning can substantially improve both task performance and output format adherence. Full fine-tuning details are provided in Appendix B.5.

Table 1: Accuracy of closed-source and open-source models on the test split of the PERCOR dataset.

Group	Model	Str Acc	PP Acc
Closed-Source	GPT-4o-mini	75.98	75.98
	GPT-4o	78.32	86.65
	GPT-4.1-nano	54.94	54.94
	GPT-4.1-mini	77.12	77.12
	GPT-4.1	88.39	88.39
	OpenAI o3	**92.18**	**92.18**
	OpenAI o4-mini	85.51	85.51
	Gemini 2.0 Flash-Lite	81.43	81.43
	Gemini 2.0 Flash	86.38	86.38
	Gemini 2.5 Flash	87.17	87.14
	Claude 3.5 Haiku	71.60	71.60
	Claude 3.7 Sonnet	91.17	91.17
Open-Source	Gemma 3n-E4B-it	59.15	59.15
	Gemma 3-1B-it	25.99	25.99
	Gemma 3-4B-it	48.32	48.32
	Gemma 3-12B-it	70.94	70.94
	Gemma 3-27B-it	76.28	76.28
	Mistral 7B Instruct v0.3	30.11	30.15
	Mistral Small 3.1 24B Instruct	68.94	68.94
	LLaMA 3.2 1B Instruct	0.79	24.12
	LLaMA 3.2 3B Instruct	25.17	25.21
	LLaMA 3.3 70B Instruct	11.23	79.56
	Aya Expanse 32B	5.85	63.27
	Command R-v01	60.0	60.0
	Command A	79.81	79.84
	Qwen 3-4B	50.33	50.33
	Qwen 3-8B	54.37	54.37
	Qwen 3-14B	69.58	69.58
	Qwen 3-30B-A3B	68.80	68.80
	Qwen 3-32B	76.54	76.54
	DeepSeek-V3	51.15	82.41
	DeepSeek-R1	**82.51**	**82.51**

4.4 Qualitative Error Analysis

Despite strong overall performance, even top models occasionally fail on examples requiring subtle syntactic, temporal, or discourse-level reasoning. Appendix Figures 14–17 present four illustrative failure cases in which the model-selected continuation is semantically implausible, grammatically incompatible, or violates the discourse relation implied by the conjunction. Each example includes a brief explanation of why the chosen option is incorrect and why the gold continuation better satisfies the coherence constraints. These qualitative observations reveal failure patterns that are not apparent from aggregate accuracy alone and highlight the value of PERCOR for probing nuanced reasoning and alignment behaviours in large language models.

4.5 Transferability to Persian QA

Prior work such as SOCIAL IQA (Sap et al., 2019) and WINOGRANDE (Sakaguchi et al., 2021) has demonstrated transferability by fine-tuning models on a commonsense benchmark and evaluating on other reasoning tasks (e.g., COPA (Gordon et al., 2011) and KNOWREF (Emami et al., 2019)). Such evaluation is not currently feasible in Persian, as PERCOR is the first structured commonsense reasoning benchmark in the language. We hope PERCOR will serve as a foundation that enables similar cross-task studies in the future. To provide an initial signal of transferability under current resource constraints, we evaluate on PQUAD (Darvishi et al., 2023), a Persian extractive QA benchmark in which many queries require implicit inference and bridging context. We fine-tune `Qwen3-32B` using LoRA with a causal language modelling objective on PERCOR, then zero-shot evaluate both the base and fine-tuned models on the PQUAD test set (964 examples) using Exact Match (EM), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002).

Fine-tuning on PERCOR yields a small improvement in EM (25.21→26.56), but substantially increases ROUGE-L (19.27→29.69, +10.4) and BLEU (16.00→23.98, +7.9). These gains suggest that the discourse-level coherence learned from PERCOR transfers to extractive QA settings where models must integrate contextual cues with implicit knowledge. We view this as an encouraging early signal of transferability and include it to underscore PERCOR's utility beyond its primary role as an evaluation benchmark.

In summary: **(i)** while PERCOR is approachable for top proprietary systems, it remains challenging for open-source models—only two models exceed 90 % accuracy, whereas the strongest open-source model still trails by approximately ten points; **(ii)** post-processing is often essential for revealing latent reasoning ability in models that express answers in free-form text rather than the required label format; **(iii)** reasoning-oriented alignment matters—`OpenAI-o3` leads all systems, and `DeepSeek-R1` surpasses `DeepSeek-V3` in strict accuracy due to stronger format adherence; **(iv)** lightweight task-specific fine-tuning remains highly effective, with small adaptations outperforming the strongest open-source zero-shot baselines; **(v)** even top-performing models exhibit systematic mistakes on examples requiring subtle syntactic, temporal, or discourse-level reasoning, underscoring PERCOR's diagnostic value; and **(vi)** the discourse-level coherence captured by PERCOR shows early signs of transferability to other Persian tasks, as reflected in improvements on PQUAD.

5 Conclusion

We introduced **PERCOR**, a 106K-example benchmark that fills a major evaluation gap for commonsense reasoning in Persian. Our conjunction-based extraction strategy yields natural sentence–completion pairs from diverse prose, and DRESS-AF selects challenging, human-written distractors while avoiding stylistic artifacts associated with LLM-generated options. Evaluating 32 models reveals a persistent ten-point gap between the strongest open-source and proprietary systems, while qualitative analysis exposes systematic failures on examples requiring subtle syntactic, temporal, or discourse-level reasoning. Together with early signs of transferability to downstream Persian tasks, these results show that PERCOR provides a robust benchmark, a fine-grained diagnostic tool, and a valuable resource for transfer learning in low-resource settings.

Future work will (i) extend our language-agnostic pipeline to other languages by adapting conjunction lists and applying DRESS-AF, and (ii) conduct expert-based human evaluation to establish a high-quality gold standard for ambiguous cases. We believe PERCOR will catalyse research on multilingual commonsense reasoning and foster the development of more robust, culturally-aware language models.

Limitations

Annotation As noted previously (see Section 4), our annotations were conducted by human annotators rather than human experts. While this approach is sufficient for broad evaluations, relying on expert annotators would likely yield more accurate and reliable assessments, particularly for complex or ambiguous cases. Moreover, we could have annotated a larger portion of the dataset to obtain a more robust and reliable estimate of human accuracy. Additionally, we could have adopted a standard annotation strategy similar to the one used in HellaSwag (Zellers et al., 2019), which involves multiple rounds of human validation and a larger set of possible answers to choose from. However, this approach requires substantially more human effort and coordination, making it more resource-intensive.

Multilingual Given that the proposed method is largely language-agnostic, we could have extended the algorithm to other languages to construct a multilingual commonsense reasoning dataset. This would have involved creating lists of conjunctions in each target language for the sentence-completion step, followed by applying the DRESS-AF algorithm accordingly.

Ethics

License In accordance with OpenAI's Terms of Use, "as between you and OpenAI... you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output"[2].

Google Gemini's terms distinguish between paid vs unpaid usage: under paid/enterprise tiers, Google does not use submitted prompts or outputs to train its models and customers retain ownership of both input and output[3]. Under unpaid or free tiers, Google may use content for product improvements, and retention policies differ.

Anthropic's Claude Terms grant users ownership of all generated outputs: "subject to your compliance with our Terms, we assign to you all of our right, title, and interest—if any—in Outputs"[4].

Based on these platform policies, we acknowledge that—under the Terms of Use for OpenAI, Google Gemini (paid/enterprise tiers), and Anthropic Claude—users retain ownership of both prompts (inputs) and generated outputs, and that the AI-produced text used in this research was obtained and employed ethically within those licensing frameworks.

Furthermore, we confirm that the outputs generated from the model were not used to train or develop models that compete with These Models. All content and model-generated assistance were applied solely for academic and illustrative purposes in the context of this research.

To generate the PerCoR dataset, we utilized textual data extracted from over 40 publicly accessible websites. All selected sources were openly available and did not impose restrictions that would preclude academic or non-commercial use.

Harmful content To curate our dataset, we selected sources with minimal sexual content and hate speech to maintain ethical standards. However, due to the complexities of open-domain language and commonsense reasoning tasks, we cannot guarantee the absence of social biases. As noted in prior work (Sakai et al., 2024; Rajani et al., 2019; Sap et al., 2020a), it remains challenging to determine when content that reflects commonsense also constitutes social bias.

[2] https://openai.com/policies/row-terms-of-use/

[3] https://ai.google.dev/gemini-api/terms

[4] https://terms.law/2024/08/24/who-owns-claudes-outputs-and-how-can-they-be-used/

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Elaf Alhazmi, Quan Z. Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *Preprint*, arXiv:2402.01512.

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: a persian natural language inference dataset. *Soft Computing*.

Anthropic. 2024. Claude3.5-haiku: fast and cost-effective reasoning model. https://www.anthropic.com/news/claude-3-5-sonnet-and-haiku. Lightweight, high-speed model outperforming Claude3 Opus on many benchmarks.

Anthropic. 2025. Claude 3.7 sonnet: Hybrid reasoning model. https://www.anthropic.com/news/

claude-3-7-sonnet. First "hybrid" model combining quick replies and extended, multi-step reasoning.

Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2021. Persianqa: a dataset for persian question answering. https://github.com/SajjjadAyobi/PersianQA.

Prajjwal Bhargava and Vincent Ng. 2022. DiscoSense: Commonsense reasoning with discourse connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10295–10310, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*.

Jiacheng Chen, Dinghan Shen, William Wang, and Diyi Zhang. 2023. Retrieval-augmented pretraining for multiple-choice question answering with knowledge triplets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1914–1929.

Shang-Hsuan Chiang, Ssu-Cheng Wang, and Yao-Chung Fan. 2022. CDGP: Automatic cloze distractor generation based on pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5835–5840, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cohere. 2024. Cohere command r: scalable 35 b instruction-following llm optimized for rag and long context. https://docs.cohere.com/docs/command-r. 35 B model with 128 K context length, multilingual and retrieval-augmented capabilities.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Naghme Jamali, Yadollah Yaghoobzadeh, and Heshaam Faili. 2022. PerCQA: Persian community question answering dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6083–6092, Marseille, France. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. 2018. Distractor generation for multiple choice questions using learning to rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mistral AI. 2025. Mistral small 3.1 (24b instruct): instruction-tuned multilingual multimodal model. `https://mistral.ai/news/mistral-small-3-1`. Advanced open-weight model with 24 B parameters, 128 K context, Apache 2.0 license.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

OpenAI. 2023. tiktoken: Tokenizer for openai models. `https://github.com/openai/tiktoken`. Accessed: 2025-05-19.

OpenAI. 2024. Gpt-4o: Openai's flagship omni-modal reasoning model. `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`. Multimodal GPT-4-level model processing text, image, and audio; claims twice the speed and half the cost of GPT-4 Turbo.

OpenAI. 2025a. Gpt-4.1 model series. `https://platform.openai.com/docs/models/gpt-4.1`. Launched April 2025; supports 1M-token context, improved instruction-following and coding performance.

OpenAI. 2025b. Openai o3 – reasoning language model. `https://platform.openai.com/docs/models/o3`. Most powerful reasoning model from OpenAI, reflecting multi-step tool use and strong performance across benchmarks.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Kai Ren and Yujie Zhu. 2021a. Knowledge-driven distractor generation for reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13453–13461.

Siyu Ren and Kenny Q. Zhu. 2021b. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *AAAI*, pages 4339–4347.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020a. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020b. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Mehran Sarmadi, Morteza Alikhani, Erfan Zinvandi, and Zahra Pourbahman. 2025. Hakim: Farsi text embedding model. *Preprint*, arXiv:2505.08435.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *Preprint*, arXiv:2409.10173.

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 Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Shuhei Watanabe. 2023. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *Preprint*, arXiv:2304.11127.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

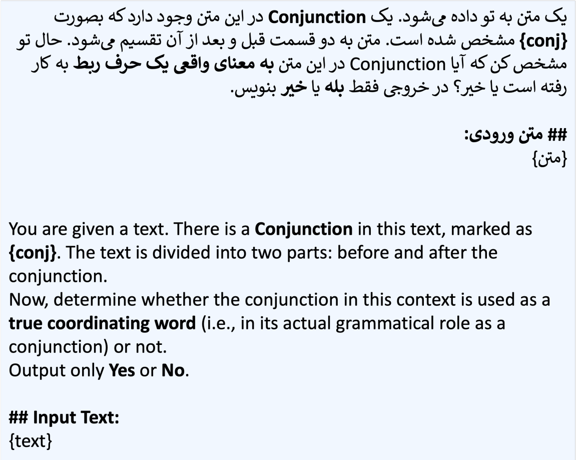
Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. Famteb: Massive text embedding benchmark in persian language. *Preprint*, arXiv:2502.11571.

A Dataset Creation

A.1 Sentence-Completion Filtering

Figure 7 presents a list of Persian conjunctions that exhibit semantic ambiguity. To address this, we employ GPT4o-mini using a binary classification prompt (shown in Figure 6) to filter out sentences in which the conjunction is used with a meaning other than the intended one. The final proportion of retained data for each connective after this filtering step is also reported in Figure 7.



یک متن به تو داده می‌شود. یک **Conjunction** در این متن وجود دارد که بصورت **{conj}** مشخص شده است. متن به دو قسمت قبل و بعد از آن تقسیم می‌شود. حال تو مشخص کن که آیا Conjunction در این متن **به معنای واقعی یک حرف ربط** به کار رفته است یا خیر؟ در خروجی فقط **بله** یا **خیر** بنویس.

## متن ورودی:
{متن}

You are given a text. There is a **Conjunction** in this text, marked as **{conj}**. The text is divided into two parts: before and after the conjunction.
Now, determine whether the conjunction in this context is used as a **true coordinating word** (i.e., in its actual grammatical role as a conjunction) or not.
Output only **Yes** or **No**.

## Input Text:
{text}

Figure 6: The prompt that we used for GPT4o-mini to detect the ambiguity of conjunctions.

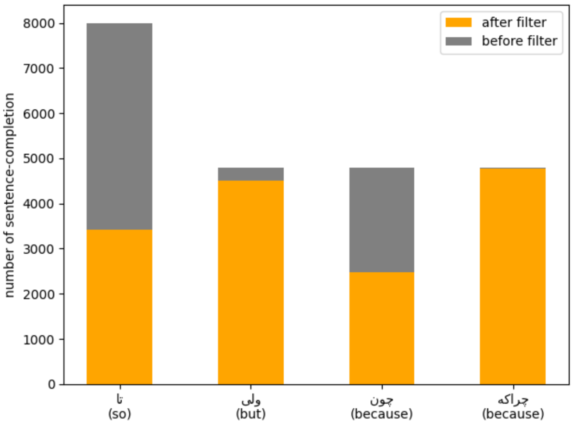


Figure 7: The number of data instances with ambiguous conjunctions before and after filtering with the GPT4o-mini model.

Furthermore, we perform an additional pass over the entire dataset using a second prompt (shown in Figure 9) to identify structurally incomplete sentence completions. At this stage, 12,117 out of 135,912 instances are flagged by the model as incomplete and removed from the dataset accordingly.

B Model Evaluation

B.1 Model Configurations

We conducted all evaluations using the vLLM inference engine for efficient serving of open-source models. Each model was run on a single NVIDIA A100 80GB GPU, except for LLaMA 3.2 70B Instruct, which required two A100 80GB GPUs due to its size. For DeepSeek variants, we used their official API endpoints, as the open-source checkpoints were not served locally.

B.2 Model Behaviour by Input Length

Owing to our conjunction-based segmentation strategy, PERCOR samples exhibit a broad range of input lengths. To assess how input length influences model performance, we analyse the correlation between sentence length and model accuracy. As shown in Figure 10, both GPT-4o-mini and Gemma 3-27B-it exhibit improved accuracy as the number of input tokens increases.

This trend suggests that longer sentence prefixes provide more contextual cues, enabling models to more reliably identify the correct completion. The result highlights a natural advantage for models when reasoning over richer, more informative contexts—an important factor to consider when designing evaluation datasets for commonsense reasoning.

B.3 Few-Shot Performance

To assess model sensitivity to minimal supervision, we conducted 1-shot and 5-shot evaluations on the PerCoR dataset using *GPT-4o-mini* and *Gemma 3-27B-it*. As shown in Figure 11, Gemma benefits modestly from few-shot prompting, improving from 76.28% (zero-shot) to 78.51% (1-shot) and 78.17% (5-shot). In contrast, GPT-4o-mini exhibits marginal or inconsistent gains, with accuracy fluctuating around its zero-shot baseline of 75.98%. These results highlight the robustness of Gemma to few-shot prompting and suggest that further gains may require stronger prompt design or fine-tuning.

B.4 Human Evaluation

We used Label Studio (Tkachenko et al., 2020-2022) to evaluate the accuracy on the test split of the dataset. Each sample was annotated independently by three human annotators. In cases where at least two annotators agreed on the same label, their consensus was taken as the final label, which was then compared with the provided label. If all

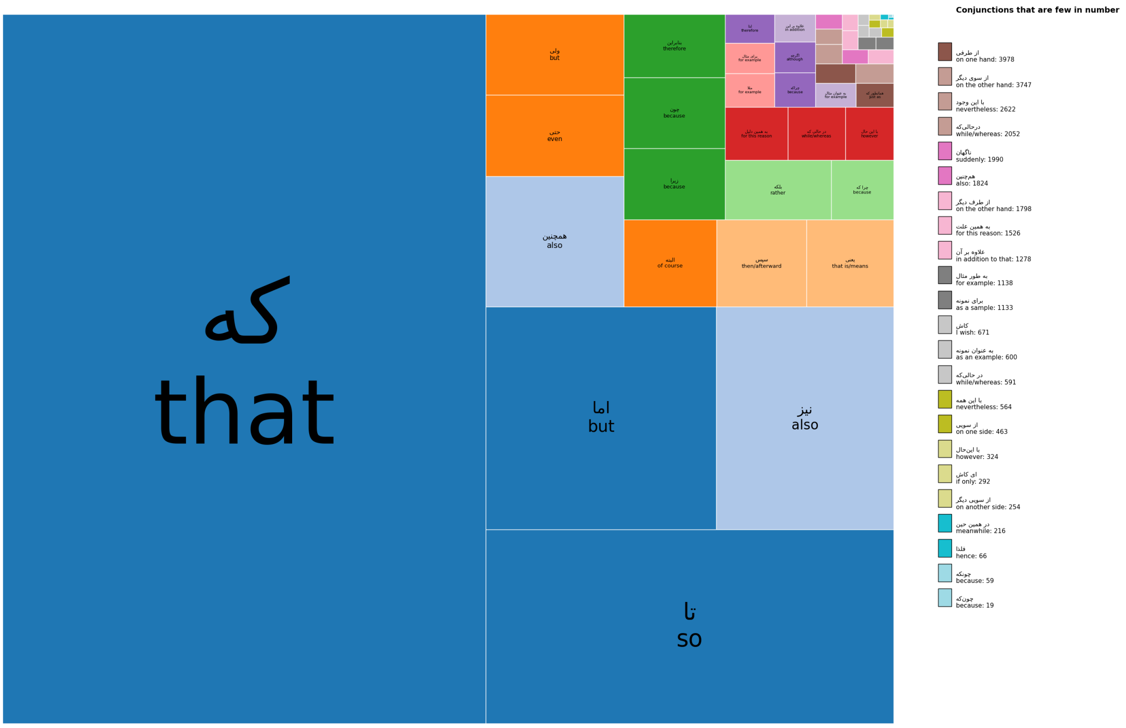
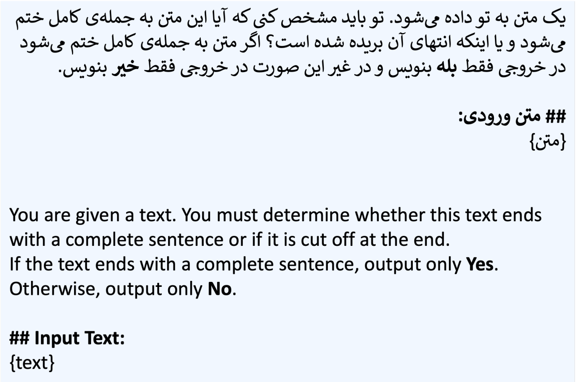


Figure 8: Treemap visualisation of conjunction words used to generate sentence–completion pairs. The area of each block corresponds to the conjunction's frequency in the dataset. Less frequently used conjunctions are shown in the adjacent panel for completeness.



یک متن به تو داده می‌شود. تو باید مشخص کنی که آیا این متن به جمله‌ی کامل ختم
می‌شود و یا اینکه انتهای آن بریده شده است؟ اگر متن به جمله‌ی کامل ختم می‌شود
در خروجی فقط **بله** بنویس و در غیر این صورت در خروجی فقط **خیر** بنویس.

## متن ورودی:
{متن}

You are given a text. You must determine whether this text ends
with a complete sentence or if it is cut off at the end.
If the text ends with a complete sentence, output only **Yes**.
Otherwise, output only **No**.

## Input Text:
{text}

Figure 9: The prompt that is leveraged by GPT4o-mini to filter out the incomplete pairs.

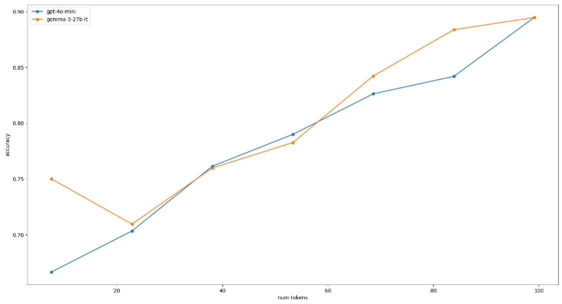


Figure 10: Accuracy of GPT-4o-mini and Gemma 3-27B-it as a function of input length. Longer prefixes tend to improve accuracy by offering more contextual information.

three annotators disagreed, the sample was considered incorrectly labelled. Importantly, annotators worked independently and were not aware of each other's selections.

B.5 Model Fine-tuning on the Dataset

To evaluate the impact of fine-tuning on PERCOR, we selected two instruction-tuned open-source models: `LLaMA-3.3-70B-Instruct` and `Qwen3-32B`. The former was quantised to 4-bit precision, while the latter was trained using `bfloat16`. We used a per-GPU batch size of 8 for `LLaMA` and 4 for `Qwen3`. Training was conducted for 2 epochs using 8×A100 80GB GPUs with DeepSpeed (Rasley et al., 2020) for distributed optimisation. We used HuggingFace (Wolf et al., 2020) for training the models.

Both models were fine-tuned using a Cosine learning rate scheduler with an initial learning rate of 5e–5 and a warmup ratio of 0.03. LoRA (Hu et al., 2022) was applied to the q, k, v, and o projection matrices within the attention layers, with hyperparameters r=4 and α=8.

Figures 12 and 13 show the training loss, evalua-

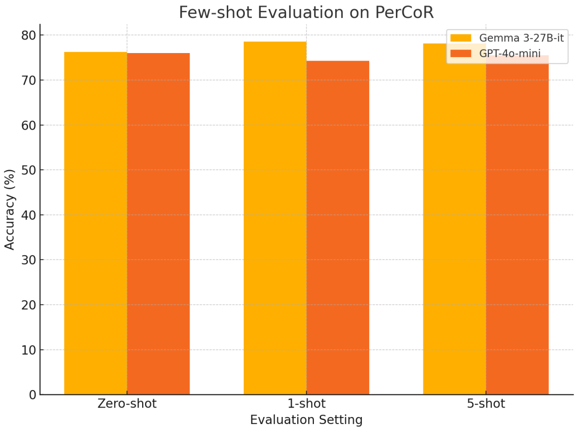


Figure 11: Accuracy of GPT-4o-mini and Gemma3-27B-it on the PERCOR dataset under zero-shot, 1-shot, and 5-shot settings.

tion loss, and evaluation accuracy over the course of training for both models. Training took approximately 2.5 hours for LLaMA3.3 and around 1 hour for Qwen3.

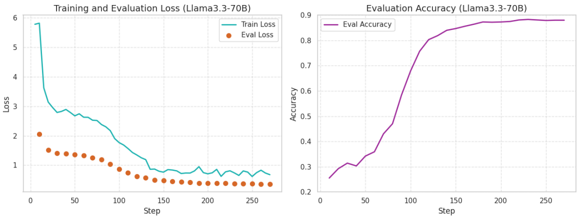


Figure 12: Training/evaluation loss and evaluation accuracy during fine-tuning of LLaMA-3.3-70B-Instruct on PERCOR.

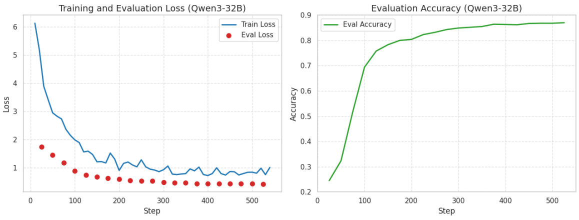
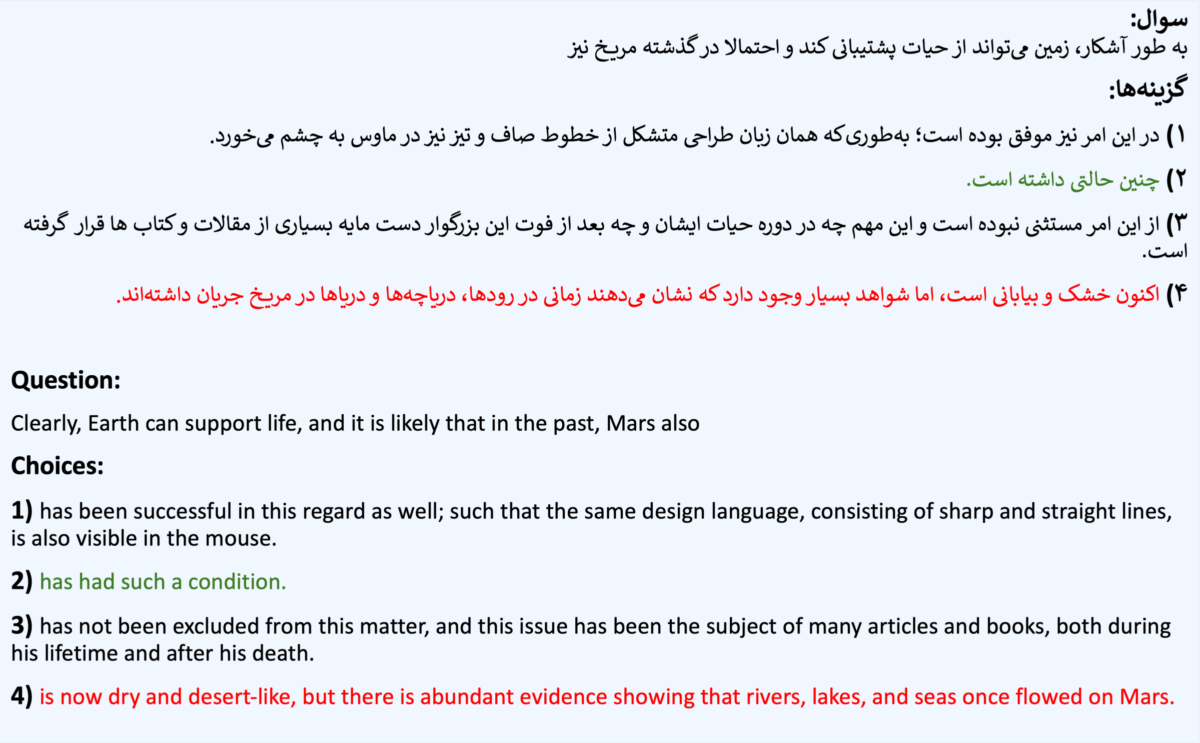


Figure 13: Training/evaluation loss and evaluation accuracy during fine-tuning of Qwen3-32B-Instruct on PERCOR.



سوال:

به طور آشکار، زمین می‌تواند از حیات پشتیبانی کند و احتمالا در گذشته مریخ نیز

گزینه‌ها:

۱) در این امر نیز موفق بوده است؛ به‌طوری‌که همان زبان طراحی متشکل از خطوط صاف و تیز نیز در ماوس به چشم می‌خورد.

۲) چنین حالتی داشته است.

۳) از این امر مستثنی نبوده است و این مهم چه در دوره حیات ایشان و چه بعد از فوت این بزرگوار دست مایه بسیاری از مقالات و کتاب ها قرار گرفته است.

۴) اکنون خشک و بیابانی است، اما شواهد بسیار وجود دارد که نشان می‌دهند زمانی در رودها، دریاچه‌ها و دریاها در مریخ جریان داشته‌اند.

Question:

Clearly, Earth can support life, and it is likely that in the past, Mars also

Choices:

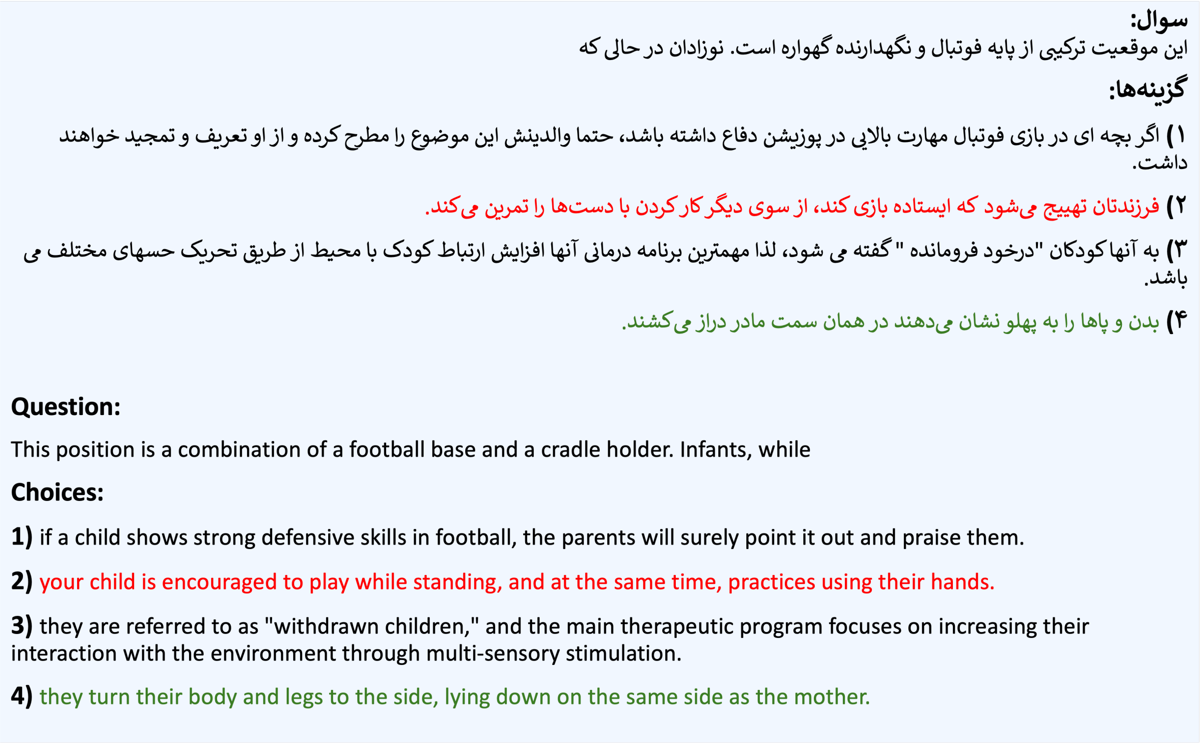
1) has been successful in this regard as well; such that the same design language, consisting of sharp and straight lines, is also visible in the mouse.

2) has had such a condition.

3) has not been excluded from this matter, and this issue has been the subject of many articles and books, both during his lifetime and after his death.

4) is now dry and desert-like, but there is abundant evidence showing that rivers, lakes, and seas once flowed on Mars.

Figure 14: An example from the PerCoR dataset where both Claude 3.7 Sonnet and GPT-4.1 incorrectly selected Option 4 (highlighted in red). While Option 4 contains true statements about the current and past state of Mars, it fails to form a coherent continuation when appended to the prompt. The question sets up a comparison referring specifically to Mars's past, expecting a grammatically and temporally consistent continuation. Option 2 correctly completes the sentence with a minimal and coherent reference to Mars's possible past habitability.



سوال:
این موقعیت ترکیبی از پایه فوتبال و نگهدارنده گهواره است. نوزادان در حالی که

گزینه‌ها:

۱) اگر بچه ای در بازی فوتبال مهارت بالایی در پوزیشن دفاع داشته باشد، حتما والدینش این موضوع را مطرح کرده و از او تعریف و تمجید خواهند داشت.

۲) فرزندتان تهییج می‌شود که ایستاده بازی کند، از سوی دیگر کار کردن با دست‌ها را تمرین می‌کند.

۳) به آنها کودکان "درخود فرومانده " گفته می شود، لذا مهمترین برنامه درمانی آنها افزایش ارتباط کودک با محیط از طریق تحریک حسهای مختلف می باشد.

۴) بدن و پاها را به پهلو نشان می‌دهند در همان سمت مادر دراز می‌کشند.

Question:

This position is a combination of a football base and a cradle holder. Infants, while

Choices:

1) if a child shows strong defensive skills in football, the parents will surely point it out and praise them.

2) your child is encouraged to play while standing, and at the same time, practices using their hands.

3) they are referred to as "withdrawn children," and the main therapeutic program focuses on increasing their interaction with the environment through multi-sensory stimulation.

4) they turn their body and legs to the side, lying down on the same side as the mother.

Figure 15: An example from the PerCoR dataset where both Claude 3.7 Sonnet and GPT-4.1 incorrectly selected Option 2 (highlighted in red). Appending Option 2 to the prompt results in a grammatically broken and incoherent sentence: "Infants, while your child is encouraged to play while standing..." — which abruptly shifts subject and verb, making no syntactic or semantic sense. The phrase "Infants, while..." requires a continuation that describes a physical or observational state of the infant. Only Option 4 satisfies this expectation with a coherent and contextually appropriate description of the infant's posture.

سوال:
به عبارت دیگر به دلیل تعطیلی بسیاری از مشاغل، درآمد آنها کاهش پیدا کرده و باعث کاهش سهم مالیات بخش حقوق از درآمد مالیاتی دولت شد. از طرف دیگر

گزینه‌ها:

۱) از محل درآمد حاصل از آن، خودروسازان بخشی از مطالبات قطعه‌سازان را بپردازند، اما این طرح اثر منفی بر بازار گذاشت، بنابراین مسکوت ماند.

۲) روند ادامه‌دار فشار هزینه‌ها همراه با نگرانی از افزایش مالیات‌ها بسیاری از کسب‌وکارها به‌ویژه در بخش خدمات را در آستانه تعطیلی قرار داده است.

۳) سهم مالیات بر ثروت نیز به دلیل پاندمی کرونا افزایش پیدا کرد.

۴) به دلیل تعطیلی گسترده ناشی از شیوع ویروس کرونا متوقف شده بودند.

Question:

In other words, due to the closure of many businesses, their income decreased, which in turn reduced the share of payroll taxes in the government's tax revenue. On the other hand

Choices:

1) from the revenue generated, automakers paid part of their debt to parts manufacturers, but this plan had a negative effect on the market and was therefore abandoned.

2) the continued pressure of rising costs, along with concerns about tax increases, has pushed many businesses—especially in the service sector—to the brink of closure.

3) the share of wealth tax also increased due to the COVID-19 pandemic.

4) they were shut down due to widespread closures caused by the spread of the coronavirus.

Figure 16: An example from the PerCoR dataset where Claude 3.7 Sonnet and GPT-4.1 both incorrectly selected Options 2 and 4, respectively. The sentence discusses how business closures led to a decline in payroll tax contributions, and the phrase "On the other hand..." introduces a contrasting development that should remain within the domain of **tax revenue**. While Option 2 is contextually plausible—highlighting economic stress—it shifts the focus away from taxation. Option 4 is even less relevant, as it redundantly repeats the cause already stated in the prompt (business closures due to COVID-19). In contrast, Option 3 presents a coherent and contrastive continuation: despite payroll tax revenue declining, the share of wealth tax increased during the pandemic. This makes Option 3 the most topically and logically aligned completion.

سوال:

گوشی X70 پرو پلاس اولین گوشی در کشور هند با تراشه اسنپدراگون ۸۸۸ پلاس محسوب می‌شود و از دوربین فوق‌العاده‌ای بهره می‌برد. با این وجود

گزینه‌ها:

۱) نسخه پرو این گوشی که کمی مقرون به صرفه‌تر به نظر می‌رسد از چیپست قدرتمند دیمنسیتی ۱۲۰۰ بهره می‌برد.

۲) همان‌طور که می‌توان حدس زد، در بازار خبری از نسخه پرو گوشی پوکو M5 نیست.

۳) سال گذشته و در پرچمداران کنونی، مایکروسافت از پیشرفته ترین چیپست کوالکام بهره گرفت و بعید به نظر می رسد که در سرفس فون، چنین کاری را انجام ندهد.

۴) این جدیدترین و بهترین تراشه‌ی پرچم‌دار کوالکام نیست، در مقایسه با تراشه‌ی اسنپدراگون G Plus 778 موجود در ناتینگ فون ۱ پیشرفتی بزرگ محسوب می‌شود.

Question:

The X70 Pro Plus is considered the first phone in India with the Snapdragon 888 Plus chip and features an exceptional camera. Nevertheless

Choices:

1) the Pro version of this phone, which seems a bit more affordable, uses the powerful Dimensity 1200 chipset.

2) as one might guess, there is no Pro version of the Poco M5 phone in the market.

3) last year and in current flagships, Microsoft used Qualcomm's most advanced chipset, and it seems unlikely that it wouldn't do the same in the Surface Phone.

4) this is not Qualcomm's newest and best flagship chip, but compared to the Snapdragon 778G Plus chip in the Nothing Phone 1, it is a significant upgrade.

Figure 17: An example from the PerCoR dataset where Claude 3.7 Sonnet incorrectly selected Option 4 as the answer. While Option 4 provides a comparison between chipsets, it fails to directly continue the original sentence, which is about the X70 Pro Plus smartphone. The phrase "Nevertheless..." sets up a contrast or qualification specifically about the phone mentioned. Option 1 correctly continues this contrast by discussing the Pro variant of the same phone and its different chipset—maintaining topical and grammatical coherence. In contrast, Option 4 shifts focus entirely to the chipset itself, breaking the discourse continuity and making it an incoherent continuation in context.