# The Sufficiency-Conciseness Trade-off in LLM Self-Explanation from an Information Bottleneck Perspective

**Ali Zahedzadeh, Behnam Bahrak**
Tehran Institute for Advanced Studies, Khatam University, Tehran, Iran
a.zahedzadeh@teias.institute, b.bahrak@teias.institute

**Abstract**

Large Language Models increasingly rely on self-explanations, such as chain of thought reasoning, to improve performance on multi step question answering. While these explanations enhance accuracy, they are often verbose and costly to generate, raising the question of how much explanation is truly necessary. In this paper, we examine the trade-off between sufficiency, defined as the ability of an explanation to justify the correct answer, and conciseness, defined as the reduction in explanation length. Building on the information bottleneck principle, we conceptualize explanations as compressed representations that retain only the information essential for producing correct answers. To operationalize this view, we introduce an evaluation pipeline that constrains explanation length and assesses sufficiency using multiple language models on the ARC Challenge dataset. To broaden the scope, we conduct experiments in both English, using the original dataset, and Persian, as a resource-limited language through translation. Our experiments show that more concise explanations often remain sufficient, preserving accuracy while substantially reducing explanation length, whereas excessive compression leads to performance degradation.

**Keywords:** large language models, self-explanation, information bottleneck, explanation sufficiency, conciseness, question answering

**Code Availability Statement.** To ensure reproducibility, all source code and experimental materials associated with this work are publicly available at: https://github.com/alizahedzadeh/LLM-Self-Explanation-IB-Bilingual

## 1. Introduction

When people answer complex exam questions, they are often expected not only to select the correct option but also to demonstrate their reasoning. Large Language Models (LLMs) are increasingly encouraged to do the same. Recent prompting strategies show that generating self-explanations most prominently chain of thought reasoning can substantially improve performance on reasoning intensive tasks (Wei et al., 2022). Instead of directly predicting an answer, the model first produces intermediate reasoning steps, which increases the likelihood of success.

Over time, several techniques have been proposed to refine this approach. Self-explanation prompting has been shown to enhance comprehension in dialogue understanding tasks (Gao et al., 2024), while models are also capable of producing spontaneous self-explanations in general reasoning scenarios (Huang et al., 2023). These advances underline the value of explanations, not only for accuracy but also for transparency and interpretability.

However, more explanation is not always better. Models frequently generate verbose, repetitive, and sometimes misleading reasoning traces. Prior studies have warned that self-explanations are not guaranteed to faithfully reflect a model's actual decision process (Madsen et al., 2024; Lyu et al., 2024). Long chains also increase latency and computational cost, while many reasoning steps are unnecessary to reach the correct answer. This raises a fundamental question: How concise can an explanation be while still sufficient to justify the answer?

From a theoretical perspective, this question connects to the Information Bottleneck (IB) framework, originally introduced by Tishby et al. (Tishby et al., 2000). The IB principle formulates learning as finding compressed representations that discard irrelevant information while preserving what is predictive of the target. This framework was later extended to deep learning (Tishby and Zaslavsky, 2015; Saxe et al., 2018) and generalized in more recent analyses that relate compression to generalization and interpretability (Kawaguchi et al., 2023; Westphal et al., 2025). From this viewpoint, a good explanation should be a minimal sufficient representation retaining just enough information to justify the prediction.

Recent work has begun to explore this balance between sufficiency and conciseness. Bassan et al. (Bassan et al., 2025) propose generating concise yet sufficient explanations through self-explaining neural architectures, while Bharti et al. (Bharti et al., 2024) and Amoukou and Brunel (Amoukou and Brunel, 2022) formalize sufficiency and necessity as distinct dimensions of interpretability. Still, a systematic, large-scale investigation of the sufficiency–conciseness trade-off in LLM-generated self-explanations is limited, particularly across different languages and reasoning settings.



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2127–2144
11–16 May 2026. 

In this work, we address this gap with an empirical study of explanation sufficiency under compression. We progressively constrain the length of explanations generated to answer questions in the ARC dataset (Clark et al., 2018), focusing on the ARC-Challenge subset, which requires multi-step reasoning beyond surface retrieval., in both English and Persian settings. Persian is included as a resource-limited language to broaden the evaluation scope. Sufficiency is assessed with a probe LLM model (Qwen 1.7B), while conciseness is measured by explanation-length reduction.

Our contributions are as follows:

- We formalize sufficiency and conciseness as complementary dimensions of explanation quality, grounded in the Information Bottleneck perspective (Tishby et al., 2000).
- We propose a general evaluation pipeline to test explanation sufficiency under progressive length constraints.
- We conduct the first bilingual study of explanation sufficiency, in English and Persian, identifying length thresholds for efficient yet sufficient self-explanations in large language models.

By quantifying the trade-off between sufficiency and conciseness, our work advances the study of explanation aware reasoning and provides practical insights for designing more efficient and trustworthy language models.

## 2. Related Work

### 2.1. Explainable AI and Attribution Methods

Early work in explainable artificial intelligence focused on interpreting predictions of black-box models through feature attributions and visualization techniques. LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) introduced post-hoc local explanation frameworks that approximate complex models with interpretable surrogates or Shapley-based attributions. Integrated Gradients (Sundararajan et al., 2017) provided a theoretically grounded method for attributing deep network predictions, while saliency maps (Simonyan et al., 2014) and attention mechanisms (Vaswani et al., 2017; Bahdanau et al., 2016; Wiegreffe and Pinter, 2019) visualized input contributions in neural models. These approaches typically rely on gradient access or model internals. In contrast, our method evaluates explanation sufficiency in a black-box setting requiring only model outputs which is crucial for evaluating proprietary large language models (LLMs) accessible solely via APIs.

### 2.2. Chain of Thought and Self-Explanations

Prompting strategies such as Chain-of-Thought (CoT) reasoning (Wei et al., 2022) and zero-shot reasoning (Kojima et al., 2022) demonstrate that eliciting intermediate reasoning steps improves performance on complex tasks. Subsequent research refined this idea through self-consistency decoding (Wang et al., 2023), least-to-most prompting (Zhou et al., 2023), self-ask decomposition (Press et al., 2023), and concise reasoning chains that perform comparably to verbose ones (Renze and Guven, 2024), which enhance reasoning reliability. Other approaches, such as ReAct prompting (Yao et al., 2023), integrate reasoning with tool use, while reflective prompting (Renze and Guven, 2025) encourages the model to critique and revise its own answers. Surveys such as (Zhang et al., 2024) summarize these developments. Collectively, these works establish that explicit reasoning boosts accuracy and interpretability but they do not address how minimal an explanation can be while remaining sufficient.

### 2.3. Faithfulness and Sufficiency of Explanations

The faithfulness of model explanations has long been a central concern. ERASER (DeYoung et al., 2020) formalized metrics such as sufficiency and comprehensiveness for evaluating rationalized NLP models, while Jacovi and Goldberg (Jacovi and Goldberg, 2020) clarified the conceptual distinction between interpretability and faithfulness. Hase and Bansal (Hase and Bansal, 2022) proposed frameworks for analyzing when models truly learn from explanations. More recent studies have extended these ideas to LLM self-explanations: Madsen et al. (Madsen et al., 2024) demonstrated that self-generated reasoning is often unfaithful to internal model behavior, and Huang et al. (Huang et al., 2023) found that explanations may be verbose, redundant, or misleading. These findings motivate direct evaluation of sufficiency whether explanations enable accurate prediction independent of surface plausibility.

### 2.4. Conciseness and Length Control

Renze and Guven (Renze and Guven, 2024) showed that concise reasoning chains can perform comparably to verbose ones, and Xu et al. (Xu et al., 2025) proposed "Chain of Draft," generating short but informative reasoning traces. Gu et al. (Gu et al., 2025) developed a black-box iterative framework for precise length control without model modification, and Wu et al. (Wu et al., 2024) observed that excessively long explanations can even

degrade reasoning quality. Despite this growing interest, no prior study has systematically quantified the sufficiency–conciseness trade-off in LLM self-explanations.

### 2.5. Information Bottleneck and Explanations

The Information Bottleneck (IB) framework (Tishby et al., 2000) views learning as compressing input representations while preserving task-relevant information. Later developments in deep learning (Tishby and Zaslavsky, 2015; Saxe et al., 2018; Kawaguchi et al., 2023; Westphal et al., 2025) extended this idea to representation learning. In the context of explainability, IB has inspired methods that treat explanations as compressed yet sufficient rationales (Paranjape et al., 2020; Li et al., 2023). Our study builds upon this perspective, viewing concise self-explanations as information-efficient justifications that retain predictive sufficiency.

### 2.6. Multilingual Self-Explanations

Recent research has explored reasoning and explanation transfer across languages. Shi et al. (Shi et al., 2022) showed that LLMs can generalize Chain-of-Thought prompting to multiple languages, while Barua et al. (Barua et al., 2025) examined multilingual reasoning with longer chains. Surveys such as Ponti et al. (Ponti et al., 2023) discuss persistent gaps in cross-lingual explanation fidelity. However, empirical work on sufficiency and conciseness in non-English contexts remains limited. Our bilingual experiments on the ARC dataset (Clark et al., 2018) fill this gap by comparing English and Persian explanations under progressive compression.

## 3. Methodology

### 3.1. Theoretical Background

Our study builds on the Information Bottleneck principle (Tishby et al., 2000), which frames learning as the problem of compressing input representations $X$ into a bottleneck variable $Z$ while preserving information relevant to the target $Y$. The general objective of the information bottleneck is

$$\max_{Z} \; I(Z;Y) - \beta I(X;Z), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information and $\beta$ is a balance parameter. In this framework, sufficiency corresponds to maximizing $I(Z;Y)$ so that the explanation $Z$ retains information required to justify the correct answer $Y$, while conciseness corresponds to minimizing $I(X;Z)$ so that the explanation does not redundantly encode irrelevant parts of the input $X$. Verbose explanations increase $I(X;Z)$ without necessarily improving $I(Z;Y)$, whereas overly short explanations risk losing sufficiency. Our approach operationalizes this trade off by constraining the length of explanations and analyzing whether sufficiency is preserved under these constraints.

Directly computing the mutual information terms in the Information Bottleneck objective is intractable for large language models. This is primarily due to their black-box nature, which prevents access to the internal probability distributions required for such calculations. Therefore, we use practical proxy metrics to approximate the trade-off. We define *Sufficiency* as the probability the model assigns to the correct answer given an explanation, and *Conciseness* as the reduction in the explanation's length. This allows us to empirically evaluate the balance between generating informative and brief explanations.

### 3.2. Constrained Explanation Generation

Let $Z$ denote the full length explanation generated for a given question. To study conciseness, we prompt the model to regenerate its explanation under explicit word length constraints. For each constraint level $v \in \{10, 20, \ldots, 90\}$, we instruct the model to produce an explanation $Z_v$ whose length is at most $(1 - v/100)$ fraction of the length of $Z$. For example, if $Z$ contains 50 words, then at the 20 percent constraint the model is required to produce an explanation of no more than 40 words. The unconstrained explanation is denoted $Z_0 := Z$. This procedure ensures that conciseness is enforced directly during generation rather than by post hoc trimming.

### 3.3. Sufficiency Evaluation

Evaluation is performed using an asymmetric setup: while multiple models generate explanations, a single fixed scorer model $M$ is employed to assess sufficiency across all conditions, ensuring comparability. For each constrained explanation $Z_v$, we construct a prompt $P_v$ comprising the question $Q$, the explanation $Z_v$, and the set of answer options $\mathcal{O} = \{A, B, C, D\}$. The scorer model outputs a probability distribution

$$p(o \mid P_v) \quad \text{for each } o \in \mathcal{O}. \tag{2}$$

Let $y$ denote the gold-standard answer. We define sufficiency as the probability assigned by the scorer to the correct answer given the explanation:

$$\text{Sufficiency}(Z_v) = p(y \mid P_v). \tag{3}$$

This metric directly quantifies the extent to which the explanation supports the correct answer. For

reference, we also evaluate a baseline prompt $P_{noexp}$ excluding any explanation. This comparison reveals the increase in the scorer's confidence attributable to explanatory content, while sufficiency itself remains defined independently as $p(y \mid P_v)$.

To prevent answer leakage, generated explanations are post-processed: any explicit mentions of answer option letters (A, B, C, D) or verbatim copies of option texts are replaced with a [MASK] symbol. This ensures that sufficiency measures genuine explanatory reasoning rather than superficial cues.

**Algorithm 1:** Computing Sufficiency with a Fixed Scorer

**Input:** Dataset $\mathcal{D}$ of items $(Q, \mathcal{O}, y, Z_v)$ with $\mathcal{O} = \{A, B, C, D\}$; fixed scorer model $M$

**Output:** Per-item sufficiency scores $\{\text{Sufficiency}(Z_v)\}_v$ and their mean

**Function** `ScoreOptions`$(M, P)$**:**

  **foreach** $o \in \mathcal{O}$ **do**

    lp$[o] \leftarrow$ sum of token log-probabilities that $M$ assigns to the string `" o"` conditioned on $P$

  **return** $\text{softmax}(\text{lp})$ ; `// distribution` $p(o \mid P)$ `over` $\mathcal{O}$

Initialize list $S \leftarrow [\,]$;

**foreach** $(Q, \mathcal{O}, y, Z_v) \in \mathcal{D}$ **do**

  $P_v \leftarrow$ prompt formed from $(Q, Z_v, \mathcal{O})$ with suffix `"The answer is "`;

  $\mathbf{p} \leftarrow$ `ScoreOptions`$(M, P_v)$ ;

  `//` $\mathbf{p}[o] = p(o \mid P_v)$

  $S$.append$(\mathbf{p}[y])$ ;

  `//` $\text{Sufficiency}(Z_v) = p(y \mid P_v)$

**return** $S$, $\dfrac{1}{|S|}\sum_{s \in S} s$;

We also report the dataset-level mean sufficiency,

$$\overline{\text{Sufficiency}} = \frac{1}{|\mathcal{D}|} \sum_{(Q,\mathcal{O},y,Z_v)\in\mathcal{D}} p(y \mid P_v).$$

### 3.4. Conciseness Measurement

Conciseness is measured by the relative reduction in explanation length enforced during generation. At each constraint level $v$, the explanation $Z_v$ is shorter by $v$% compared to the unconstrained explanation $Z_0$. By evaluating sufficiency across all levels, we can empirically analyze the trade-off between conciseness and sufficiency predicted by the information bottleneck framework.

### 3.5. Pipeline Summary

The complete pipeline consists of three stages as illustrated in Figure 1. First, explanations are generated by multiple models under a shared prompting scheme. Second, constrained versions of each explanation are generated at predefined reduction levels through explicit length control in the prompt. Third, the fixed scorer model evaluates all explanations and returns sufficiency values defined as probabilities assigned to the correct answer. Baseline runs without explanations are included for comparison. Throughout the pipeline, structured logging and storage in the Comma-Separated Values (CSV) format ensure reproducibility and enable systematic quantitative analysis.

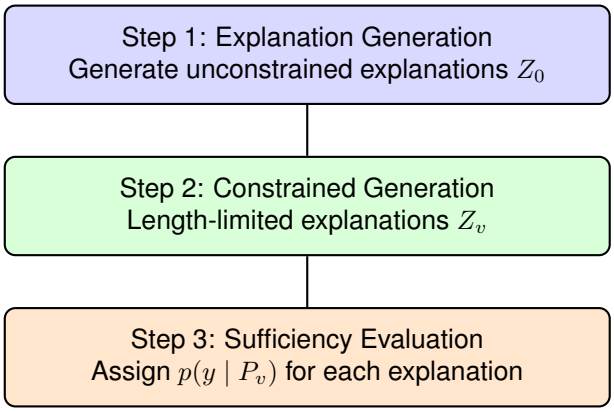


Figure 1: Pipeline of the proposed methodology. Explanations are generated by LLMs, then regenerated under explicit length constraints, and finally evaluated for sufficiency using a fixed scorer model.

## 4. Experimental Setup

### 4.1. Dataset

We conducted all experiments on the ARC Challenge dataset (Clark et al., 2018), a benchmark of multiple choice science questions designed to test advanced reasoning capabilities. The dataset contains 7,787 questions in total, divided into an Easy Set and a Challenge Set. The Challenge Set, which we use in this study, consists of 2590 four choice questions that are known to be difficult for retrieval based and co-occurrence based algorithms. Each question requires non-trivial reasoning rather than surface level matching.

To broaden the scope of evaluation, we also create a Persian version of the dataset by translating the original English questions and answer options. Persian is included as a resource-limited language to evaluate the robustness of explanation sufficiency across linguistic settings.

### 4.2. Models and Prompting

We evaluated seven large language models from diverse providers as explanation generators: GPT-4o-mini (Hurst and et al., 2024) (OpenAI), Claude 3 Haiku (Anthropic, 2024) (Anthropic), Llama 4 Scout (Meta AI, 2025) (Meta, 109 billion total parameters with 17 billion active in MoE architecture (Sanseviero et al., 2023; Mu and Lin, 2025)), Gemini 2.0 Flash (Gemini Team, 2024) (Google), Cohere Command R (Cohere, 2024) (Cohere, 35 billion parameters), DeepSeek-V3.1 (DeepSeek-AI, 2025) (DeepSeek, 671 billion total parameters with 37 billion active in MoE architecture (Sanseviero et al., 2023; Mu and Lin, 2025)), and Mistral Small 3.1 (Jiang and et al., 2024) (Mistral, 24 billion parameters). These models were selected for their cost-effectiveness, as the evaluation required generating outputs in multiple length variants, which would have incurred substantial expenses if more advanced models had been employed. The variation in parameter counts, including dense and Mixture-of-Experts (MoE) architectures, further supports efficient inference while maintaining diverse capabilities. All models were accessed through the unified OpenRouter API (OpenRouter AI, 2025), which standardizes query parameters and temperature control across providers. A single fixed scorer model, Qwen3 1.7B (Yang and et al., 2025), was used to evaluate explanation sufficiency, ensuring consistent judgments across all generated explanations. All models are prompted with a uniform template. Each model is instructed to produce a final prediction in the form of one option among A, B, C, or D, and to accompany it with a natural language explanation. For length constrained variants, models are instructed to regenerate their explanations under explicit word length limits defined as percentages of the original explanation length. Prompt templates are provided in the final version of this paper.

### 4.3. Implementation Details

We evaluate on the full 2590 question ARC Challenge dataset. For each generated explanation, we create multiple constrained versions corresponding to nine reduction levels ranging from 10 percent to 90 percent of the original explanation's length. To prevent answer leakage, we apply masking to all explanations prior to evaluation. If an explanation explicitly mentions option labels (A, B, C, D) or copies answer text verbatim, such tokens are replaced with a [MASK] symbol. The scorer model then evaluates sufficiency for each explanation by assigning probabilities to the four answer options. All runs are executed in Python with structured logging and storage in CSV format to facilitate reproducibility.

## 5. Results

In this section, we present our experimental findings on the ARC-Challenge dataset in both English and Persian. We evaluate the quality of model-generated explanations under progressively tighter length constraints using four complementary metrics that jointly capture task performance, explanatory sufficiency, and semantic preservation:

1. **Accuracy:** the proportion of correctly predicted answers produced by the model. This metric reflects the overall task performance and serves as a direct measure of reasoning success.

2. **Sufficiency:** the probability assigned by the fixed scorer model to the correct answer, given the explanation. This metric quantifies how effectively an explanation enables the correct prediction, independent of surface plausibility.

3. **Embedding Similarity:** the cosine similarity between sentence-level embeddings of the base explanation and its compressed variant. This measures semantic alignment and helps determine whether shorter explanations retain the same meaning.

Together, these three perspectives allow us to systematically investigate the trade-off between conciseness and explanatory adequacy.

### 5.1. Baseline Performance of the Scorer Model

Before analyzing the effect of explanations on sufficiency, we first report the *baseline performance* of the scorer model itself. In this setting, the model only receives the question and the answer options, without any explanation. We measure two metrics:

- **Baseline Accuracy:** the percentage of cases where the selected option by the scorer model matches the gold answer.

- **Baseline Sufficiency:** the average probability assigned to the gold option by the scorer model in the absence of explanations.

This baseline provides a reference point for interpreting subsequent improvements or declines when explanations are added or shortened. The results are presented in Table 1.

Table 1: Baseline accuracy (%) and sufficiency (%) of the scorer model without explanations.

| Language | Accuracy | Sufficiency |
|---|---|---|
| English | 71.17 | 80.71 |
| Persian | 47.85 | 72.82 |

### 5.2. Embedding Similarity Analysis

Beyond accuracy and sufficiency, we further analyze the *semantic stability* of explanations under progressive compression. This analysis measures how closely shortened explanations preserve the original meaning, thereby quantifying the extent of semantic drift caused by conciseness constraints. For each question, we encode both the base explanation $Z_0$ and its length-limited variants ($Z_{10}$–$Z_{90}$) using the `Qwen/Qwen3-Embedding-0.6B` model. We then compute the cosine similarity between the embedding vectors of each pair and average the results over the dataset. The resulting score lies in the range $[0, 1]$, where higher values indicate stronger preservation of meaning. To visualize cross-model and cross-language patterns, we report the mean similarity across all constraint levels as a heatmap (Figure 2). Each cell represents the average embedding similarity between the unconstrained explanation and its compressed counterpart for a specific model and constraint level. Darker shades correspond to higher semantic overlap. Overall, the heatmaps reveal a consistent downward gradient from left to right, indicating semantic drift under compression. However, the rate of degradation varies across models and languages. In English, models such as Claude 3 Haiku and Gemini 2.0 Flash maintain relatively high similarity scores (above 0.80) at moderate compression levels, suggesting that their explanations are more information-dense. In contrast, Persian explanations degrade more sharply after the 40% level, reflecting both linguistic complexity and lower model exposure to Persian training data. Larger models like DeepSeek-V3.1 (671 billion total parameters) exhibit sustained similarity above 0.80 in English, potentially due to greater information redundancy. Interestingly, semantic similarity does not always align with sufficiency: even when explanations remain semantically close to the original ($\text{sim} \geq 0.85$), their sufficiency scores can drop noticeably. This observation indicates that beyond lexical or semantic preservation, structural and causal cues in the explanation are also essential for maintaining reasoning effectiveness.

### 5.3. Overall Accuracy and Sufficiency

Table 2 reports accuracy and sufficiency scores for all seven LLMs using full explanations ($Z_0$) on the ARC Challenge dataset. Results are shown for both the original English setting and the Persian translated version. Across all models, explanations substantially improve performance compared to the no-explanation baseline (presented in Table 1), confirming that self-explanations provide critical support for reasoning. However, we also observe a consistent gap between English and Persian performance, reflecting the increased difficulty of reasoning in a resource-limited language. As seen in Table 2, all models achieve higher accuracy and sufficiency with explanations, with DeepSeek V3.1 and Mistral Small 3.1 performing strongest in English, while GPT-4o-mini and Gemini 2.0 Flash lead in Persian. Persian results consistently lag behind English, highlighting the challenges of applying explanation-based reasoning in low-resource linguistic settings. Notably, larger models such as DeepSeek-V3.1 (671 billion total parameters) excel in English, whereas smaller ones like GPT-4o-mini (8 billion parameters) demonstrate resilience in Persian, suggesting that parameter scale interacts with language-specific training.

### 5.4. Effect of Conciseness on Accuracy and Sufficiency

We next analyze how progressively shortening explanations impacts both task accuracy and sufficiency. Figures 3 and 4 present results across all generator models on the ARC Challenge dataset in Persian and English, respectively. Several consistent patterns emerge. First, both accuracy and sufficiency decrease as explanations are shortened, confirming that longer explanations generally provide more reliable reasoning support. Second, the degradation is steeper in Persian compared to English, reflecting the additional difficulty of generating effective concise explanations in a low-resource language. Third, model differences are evident: in English, DeepSeek V3.1 and GPT-4o-mini demonstrate greater robustness, maintaining accuracy scores above 0.84 even at 90% constraint levels, whereas LLaMA 4 Scout and Mistral Small 3.1 exhibit sharper declines, dropping below 0.75. In Persian, Gemini 2.0 Flash and Claude 3 Haiku preserve performance more effectively under high compression, with sufficiency scores remaining above 0.66 at 90% constraints, compared to steeper drops in models like Cohere Command R and Mistral Small 3.1. These variations suggest that differences in instruction-following capabilities, multilingual training, and alignment may contribute to enhanced robustness under conciseness constraints.

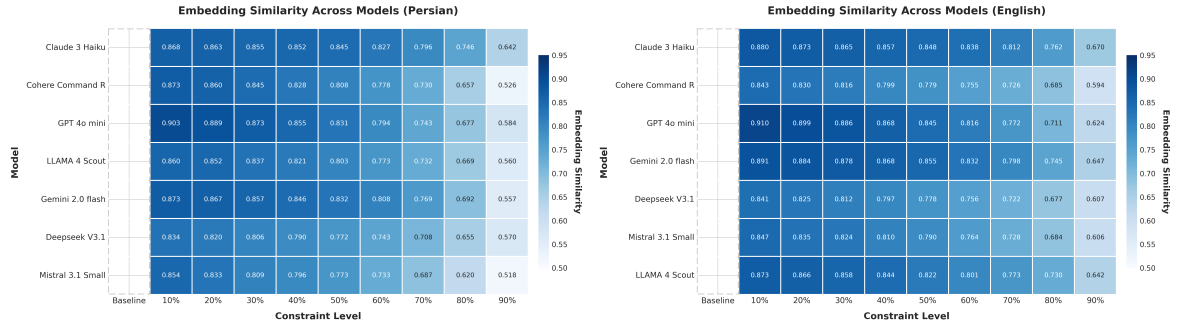


Figure 2: Average embedding similarity between base and length-constrained explanations across models and constraint levels, shown as heatmaps for Persian (left) and English (right). Darker shades indicate stronger semantic preservation.

Table 2: Accuracy and sufficiency (%) with full explanations ($Z_0$) for all models on ARC Challenge in English and Persian.

| Model | English Acc | English Suff | Persian Acc | Persian Suff |
|---|---|---|---|---|
| GPT-4o-mini | 89.34 | 86.18 | **82.60** | **79.18** |
| Claude 3 Haiku | 86.20 | 83.05 | 78.38 | 74.77 |
| LLaMA 4 Scout | 89.76 | 86.76 | 79.62 | 75.81 |
| Gemini 2.0 Flash | 89.77 | 85.74 | 82.06 | 78.08 |
| Cohere Command R | 85.00 | 81.81 | 69.23 | 66.12 |
| DeepSeek V3.1 | **90.74** | **87.32** | 80.24 | 76.60 |
| Mistral Small 3.1 | 90.00 | 86.18 | 75.04 | 71.63 |

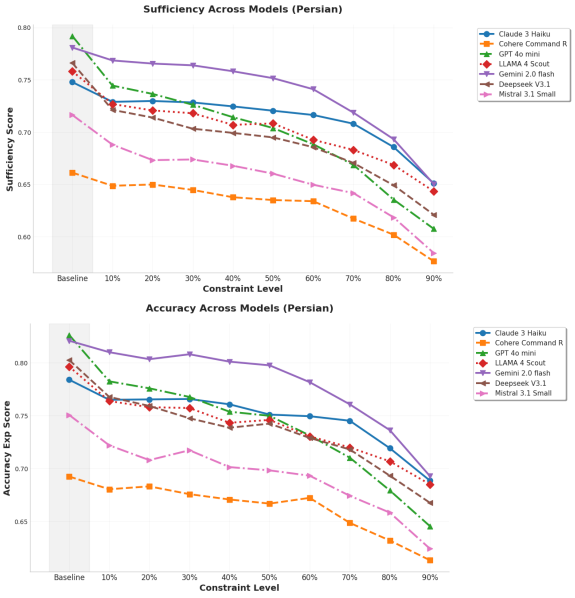


Figure 3: Sufficiency (up) and Accuracy (down) across explanation length constraints in Persian.

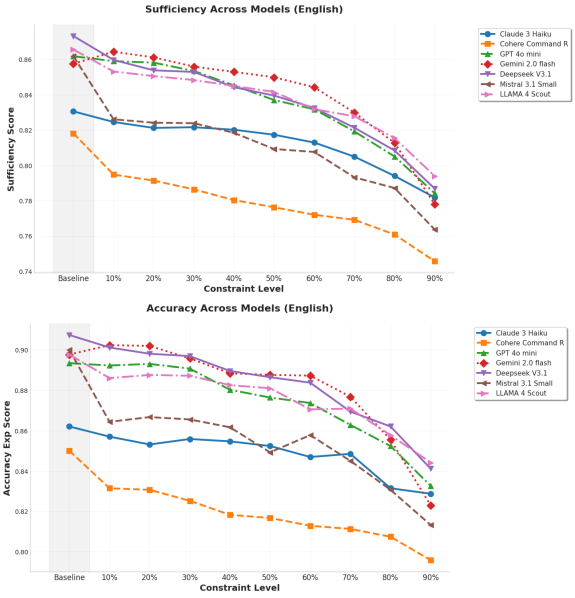


Figure 4: Sufficiency (up) and Accuracy (down) across explanation length constraints in English.

## 6. Conclusion

Our work examined the trade-off between explanation sufficiency and conciseness in large language models through the lens of the Information Bottleneck principle. We proposed a systematic evaluation pipeline to constrain and assess self-explanations, applying it to both English and Persian versions of the ARC-Challenge dataset. Our findings reveal that explanations can often be substantially shortened while preserving sufficiency, indicating that many reasoning chains contain re-

dundant steps. This suggests that concise explanations not only reduce computational cost but also align more closely with human-like efficiency in reasoning. Future work will extend the framework to open-ended tasks, investigate automatic conciseness control, and explore sufficiency preservation under multilingual and multimodal conditions.

## 7. Ethical Considerations

All datasets and models employed in this research are publicly available and adhere to established ethical guidelines for AI research. The ARC-Challenge dataset comprises 2,590 genuine grade-school-level, multiple-choice science questions designed to advance question-answering methodologies, with no inclusion of personal, sensitive, or identifiable information. The Persian translation of the dataset was generated automatically and manually reviewed for fidelity, ensuring it contains no human-identifiable content or biases that could arise from mistranslation. All large language models were accessed via official APIs in full compliance with their respective usage policies, mitigating risks associated with unauthorized data handling. No human subjects participated in this study, thereby eliminating concerns related to informed consent or privacy. Additionally, we acknowledge potential environmental impacts from LLM inference and advocate for energy-efficient practices in future extensions of this work.

## References


Salim I. Amoukou and Nicolas Brunel. 2022. Consistent sufficient explanations and minimal local rules for explaining the decision of any classifier or regressor. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, pages 8027–8040.

Anthropic. 2024. Claude 3.5 Sonnet Model Card Addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Arpan Barua, Arnav Singh, and Zhou Denny. 2025. Long chain-of-thought reasoning in multilingual large language models.

Shahaf Bassan, Ron Eliav, and Shlomit Gur. 2025. Explain yourself, briefly! self-explaining neural networks with concise sufficient reasons. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Beepul Bharti, Paul Yi, and Jeremias Sulam. 2024. Sufficient and necessary explanations (and what lies in between).

Cohere. 2024. Introducing command r: A new era of scalable ai for enterprises. Web page.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458. Association for Computational Linguistics.

Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, Fei Huang, and Yongbin Li. 2024. Self-explanation prompting improves dialogue understanding in large language models. In *Proceedings of LREC-COLING 2024*, Torino, Italy. ELRA and ICCL.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, Kun Zhu, Lei Huang, Ting Liu, Bing Qin, and Tat-Seng Chua. 2025. Length controlled generation for black-box LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16878–16895, Vienna, Austria. Association for Computational Linguistics.

Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, pages 29–39. Association for Computational Linguistics.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations.

Aaron Hurst and et al. 2024. GPT-4o System Card.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.

Albert Q. Jiang and et al. 2024. Mixtral of Experts.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023. How does information bottleneck help deep learning?

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA. Curran Associates Inc.

Qintong Li, Zhiyong Wu, Lingpeng Kong, and Wei Bi. 2023. Explanation regeneration via information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12081–12102. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, pages 4768–4777, Long Beach, CA, USA. Curran Associates Inc.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

Siyuan Mu and Sen Lin. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications.

OpenRouter AI. 2025. Openrouter: Unified api access for large language models. https://openrouter.ai/docs/api-reference/overview. Accessed: 2025-10-16.

Bhargavi Paranjape, Jing Chen, Graham Neubig, and Zachary C. Lipton. 2020. Information bottleneck for interpretable and generalizable sentence embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8703–8717. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Ivan Vulic, Roi Reichart, and Sebastian Ruder. 2023. Cross-lingual generalization in large language models: A survey.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Matthew Renze and Erhan Guven. 2024. The benefits of a concise chain of thought on problem-solving in large language models.

Matthew Renze and Erhan Guven. 2025. Self-reflection in llm agents: Effects on problem-solving performance.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco, California, USA. Association for Computing Machinery.

Omar Sanseviero, Lewis Tunstall, Philipp Schmid, Sourab Mangrulkar, Younes Belkada, and Pedro Cuenca. 2023. Mixture of experts explained.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. 2018. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations (ICLR)*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 3319–3328, Sydney, Australia. JMLR.org.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)*, Red Hook, NY, USA. Curran Associates Inc.

Charles Westphal, Stephen Hailes, and Mirco Musolesi. 2025. A generalized information bottleneck theory of deep learning.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Zhaoyang Wu, Xingxuan Zhang, Qiang Li, Guanhua Chen, and Zhiyuan Liu. 2024. More is less: When more explanations hurt reasoning in large language models.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less.

An Yang and et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Yang, Karthik Narasimhan, Yuan Cao, Dawei Song, Wenhao Li, and Nan Du. 2023. React: Synergizing reasoning and acting in language models. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Zhexin Zhang, Weize Zhao, Ruichao Xu, Zhenyu Li, Qi Zhu, Hao Zhou, and Jingjing Zhang. 2024. Chain-of-thought reasoning in large language models: A survey.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.

## 8. Language Resource References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana. AI2 Reasoning Challenge (ARC) Dataset, ARC-Challenge subset.

# A. Model Specifications and Computational Infrastructure

To ensure full replicability and address concerns regarding potential evaluation bias from a single scorer, we provide comprehensive details of the models and hardware used in our study. While the primary experiments were conducted using `Qwen3-1.7B`, we introduced two additional open-source scorers to verify that the observed sufficiency–conciseness trade-off is stable across different architectures.

## 1.1. Model Selection and Roles

We utilize a diverse set of models categorized by their specific roles in the study. All local evaluation and embedding models are hosted on our internal GPU infrastructure.

Table 3: Summary of models used for inference, evaluation, and semantic analysis. All generator models were utilized to produce both base ($Z_0$) and concise ($Z_v$) versions.

| Model Name | Provider | Access Type | Role in Study |
|---|---|---|---|
| GPT-4o-mini | OpenAI | API | Base & Concise Generation |
| Claude 3 Haiku | Anthropic | API | Base & Concise Generation |
| Llama 4 Scout | Meta | API | Base & Concise Generation |
| Gemini 2.0 Flash | Google | API | Base & Concise Generation |
| Cohere Command R | Cohere | API | Base & Concise Generation |
| DeepSeek-V3.1 | DeepSeek | API | Base & Concise Generation |
| Mistral Small 3.1 | Mistral | API | Base & Concise Generation |
| Qwen3-1.7B | Alibaba | Local | Primary Scorer (Sufficiency) |
| Gemma-3-4B-PT | Google | Local | Stability Validation Scorer |
| Llama-3.2-3B | Meta | Local | Stability Validation Scorer |
| Qwen3-Embed-0.6B | Alibaba | Local | Embedding (Similarity Check) |

## 1.2. Experimental Stability and Multi-Scorer Validation

To ensure the robustness of our results, we conducted a cross-model stability analysis. The consistency of sufficiency scores across the different model families (Qwen, Gemma, and Llama) confirms that the performance degradation observed under extreme conciseness is a generalizable phenomenon.

## 1.3. Hardware and Hyperparameters

All local evaluation tasks were executed on a GPU server to maintain consistent environmental conditions.

- **GPU Hardware**: NVIDIA Quadro RTX 6000 with 24 Gb of VRAM.
- **Inference Settings**: For consistency across all model types, the *temperature* was set to 0.3.
- **Token Limits**: The *max_tokens* parameter was set to `None` to allow models to generate complete explanations without truncation.
- **Environment**: Local models were run using the `transformers` library in a Linux-based environment.

# B. Prompt Templates

This appendix provides the complete set of prompt templates utilized across both English and Persian datasets to ensure structural consistency and facilitate the Information Bottleneck (IB) operations. These templates are designed to maximize replicability in accordance with reviewer feedback.

## 2.1. System Role and Universal Constraints

A foundational system prompt was utilized for all seven inference models to enforce strict output formatting.

**System Prompt**

```
You are a reasoning assistant for multiple-choice QA.
Always respond in this exact format:

<prediction>[A/B/C/D]</prediction>
<explanation>[Your short explanation]</explanation>
```

## 2.2. Base Generation ($Z_0$) Templates

The following templates were used for the initial elicitation of unconstrained self-explanations. The Persian template was specifically crafted to ensure the model responds in Persian while maintaining the same logical constraints as the English version.

**English Base Template**

```
You are given a multiple-choice question.

Step 1: Based on your knowledge and reasoning, select the most likely correct answer.
Step 2: Justify your answer with clear reasoning and explanation.

Instructions:
- Use logical reasoning to determine the best answer.
- Do not reference the other answer options in your explanation.
- Keep the explanation concise but informative (2-4 sentences).

---
Question: {question}
Options: A) {option_A} | B) {option_B} | C) {option_C} | D) {option_D}

Respond in this format:
<prediction>[A/B/C/D]</prediction>
<explanation>[Your reasoning and justification for the answer]</explanation>
```

**Persian Base Template**

```
You are given a multiple-choice question written in **Persian**.

Step 1: Based on your knowledge and reasoning, select the most likely correct answer.
Step 2: Justify your answer with clear reasoning and explanation.

Instructions:
- Use logical reasoning to determine the best answer.
- Do not reference the other answer options in your explanation.
- Read the question and options carefully (they are in Persian).
- Choose the most likely correct option (A, B, C, or D).
- Keep the explanation concise but informative in **Persian language** (2-4 sentences).
- Provide clear reasoning for your choice.

---
Question: {question}
Options: A) {option_A} | B) {option_B} | C) {option_C} | D) {option_D}

Respond in this format:
<prediction>[A/B/C/D]</prediction>
<explanation>[Your explanation in Persian]</explanation>
```

## 2.3. Information Bottleneck: Concise Rewrite Prompt

To implement the conciseness-sufficiency trade-off, we utilized a targeted rewrite prompt. This phase enforces a strict word count ($target\_words \pm 2$) and prohibits the use of option labels or direct keyword copying to ensure the scorer relies on reasoning rather than surface-level patterns.

**IB Concise Rewrite Template**

```
You are given a multiple-choice question, its options, and a previous explanation.

Your task:
- Keep the predicted answer fixed: {prediction}.
- Rewrite the explanation in a shorter, more concise version.
- Limit the rewritten explanation to about {target_words} words (+/- 2 words).
- Do NOT mention option letters (A/B/C/D) in the explanation.
- Do NOT copy option texts or keywords directly.
- Focus only on the reasoning behind the correct choice.

Question:
{question}

Options:
A) {choices_A}
B) {choices_B}
C) {choices_C}
D) {choices_D}

Original Explanation:
{original_explanation}

Output Format:
<answer>{prediction}</answer>
<concise_explanation>
[Rewritten concise reasoning here]
</concise_explanation>
```

### 2.4. Scorer Evaluation Templates

The following standardized formats were used by the local scorers (*Qwen-3*, *Gemma-3*, and *Llama-3.2*) to extract log-probabilities for the answer options across all retention levels ($Z_{10}$–$Z_{90}$).

- **Baseline (No Explanation)**:
  `{Question} \n Options: \n A) {A} \n B) {B} ... \n The answer is`
- **Explanation-Augmented ($Z_v$)**:
  `{Question} \n Explanation: {Explanation} \n Options: \n A) {A} \n B) {B} ... \n The answer is`

## C. Comprehensive Qualitative Examples and Error Analysis

This section provides detailed qualitative analyses, metadata, and probability traces for selected examples from both the Persian and English datasets. All generation results reported in this section are produced by the inference model **GPT-4o-mini**, while explanation quality and sufficiency judgments are evaluated using the **Qwen3-1.7B** evaluator model.

We categorize the examples into three representative behavioral patterns: (1) *Self-Correction via Explanation*, where the model revises an initially incorrect baseline prediction after generating a full explanation; (2) *Explanation-Induced Error*, where reasoning leads the model to deviate from an originally correct prediction; and (3) *Robust Reasoning Under Compression*, where the model preserves the correct answer under both full and shortened explanations.

These cases highlight how explanation length and structure interact with parametric biases and confidence calibration. In particular, some examples demonstrate a failure to reach the predefined sufficiency threshold: although the explanation increases the probability assigned to the correct option, the model remains indecisive due to strong lexical or thematic priors (e.g., ecological keywords).

### 3.1. Example: Explanation-Induced Correction (Persian)

Table 4: Persian Dataset Example: Full Explanation Corrects the Model.

| Field | Content / Value |
|---|---|
| **Question** | کدام گزینه با هدف یک جامعه سبز در تضاد است؟ |
| **Options** | با چوب منازل خود را گرم می‌کنند؛ سبزیجات را قبل از پختن می‌شویند؛ با خودروهای با سوخت بنزینی رانندگی می‌کنند؛ در باغ جامعه گل می‌کارند |
| **Correct Answer** | C |
| **Baseline Prediction (No Explanation)** | D \| Prob: 0.7101 |
| **Full Explanation** | در جوامع سبز هدف کاهش استفاده از منابع غیرقابل تجدید است. استفاده از خودروهای با سوخت بنزینی باعث افزایش مصرف انرژی‌های فسیلی و آلودگی می‌شود. |
| **Prediction (With Full Explanation)** | C \| Prob: 0.9526 |

### 3.2. Example: Explanation-Induced Correction (English)

Table 5: English Dataset Example: Full Explanation Corrects the Model.

| Field | Content / Value |
|---|---|
| **Question** | Which part of an atom occupies the majority of the volume? |
| **Options** | nucleus; protons; neutrons; electrons |
| **Correct Answer** | D |
| **Baseline Prediction (No Explanation)** | A \| Prob: 0.8319 |
| **Full Explanation** | The majority of an atom's volume is occupied by the electron cloud, which surrounds the nucleus. Electrons exist in orbitals extending far from the nucleus. While the nucleus contains most of the mass, electrons account for most of the volume. |
| **Prediction (With Full Explanation)** | D \| Prob: 0.8508 |

### 3.3. Example: Performance Degradation Under Short Explanation (Persian)

Table 6: Persian Dataset Example: Shortened Explanation Causes Incorrect Prediction.

| Field | Content / Value |
|---|---|
| **Question** | کدام گزینه با هدف یک جامعه سبز در تضاد است؟ |
| **Options** | با چوب منازل خود را گرم می‌کنند؛ سبزیجات را قبل از پختن می‌شویند؛ با خودروهای با سوخت بنزینی رانندگی می‌کنند؛ در جامعه گل می‌کارند |
| **Correct Answer** | C |
| **Baseline Prediction (No Explanation)** | D \| Prob: 0.7101 |
| **Short Explanation** | هدف جامعه‌های سبز کاهش مصرف منابع غیرقابل تجدید است. |
| **Prediction (With Short Explanation)** | A \| Prob: 0.7153 |

### 3.4. Example: Failure Under Explanation Compression (English)

Table 7: English Dataset Example: Shortened Explanation Fails to Correct the Error.

| Field | Content / Value |
|---|---|
| **Question** | Which part of an atom occupies the majority of the volume? |
| **Options** | nucleus; protons; neutrons; electrons |
| **Correct Answer** | D |
| **Baseline Prediction (No Explanation)** | A \| Prob: 0.8319 |
| **Short Explanation** | Electrons occupy most of the atom's volume. |
| **Prediction (With Short Explanation)** | A \| Prob: 0.6184 |

### 3.5. Example: Robust Performance Across Explanation Variants (Persian)

Table 8: Persian Dataset Example: Model Remains Correct Under Full and Short Explanations.

| Field | Content / Value |
|---|---|
| **Question** | سندرم تونل کارپ می‌تواند ناشی از حرکات تکراری مچ دست باشد. این نمونه‌ای از استفاده از فناوری با وجود چه چیزی است؟ |
| **Options** | کمک به جامعه؛ تأثیر منفی؛ پیشرفت‌های پزشکی؛ ویژگی‌هایرِنِنت |
| **Correct Answer** | B |
| **Baseline Prediction (No Explanation)** | B \| Prob: 0.5062 |
| **Full Explanation** | سندرم تونل کارپال ناشی از حرکات تکراری مچ و دست‌ها است که می‌تواند به دلیل استفاده مداوم از تکنولوژی مانند تایپ کردن روی کیبورد ایجاد شود. این وضعیت نشان‌دهنده تأثیر منفی استفاده از تکنولوژی است. |
| **Prediction (With Full Explanation)** | B \| Prob: 0.9982 |
| **Short Explanation** | حرکات تکراری باعث آسیب می‌شوند. |
| **Prediction (With Short Explanation)** | B \| Prob: 0.9936 |

### 3.6. Example: Robust Performance Across Explanation Variants (English)

Table 9: English Dataset Example: Model Remains Correct Under Full and Short Explanations.

| Field | Content / Value |
|---|---|
| **Question** | The snowshoe hare sheds its fur twice a year. In summer, its fur is brown; in winter, it turns white. Which statement best explains the advantage of shedding fur? |
| **Options** | Shedding fur keeps the hare clean; Shedding fur helps the hare move quickly; Shedding fur keeps the hare's home warm; Shedding fur helps the hare blend into its habitat |
| **Correct Answer** | D |
| **Baseline Prediction (No Explanation)** | D \| Prob: 0.7264 |
| **Full Explanation** | Shedding fur allows the snowshoe hare to adapt its coat color to seasonal changes, providing camouflage against predators. Brown fur blends with the forest floor in summer, while white fur blends with snow in winter, improving survival. |
| **Prediction (With Full Explanation)** | D \| Prob: 0.9687 |
| **Short Explanation** | Adapting coat color enhances camouflage. |
| **Prediction (With Short Explanation)** | D \| Prob: 0.9446 |

## D. Comprehensive Statistical Analysis: English vs. Persian

This section provides the complete statistical breakdown of the Information Bottleneck operations. We report results for all nine retention levels ($Z_{10}$ to $Z_{90}$), where $Z_v$ represents the percentage of tokens retained (e.g., $Z_{10}$ indicates 90% compression). All tests were performed on a sample size of $N = 20,647$ for English and Persian.

Table 10: Comparative Statistical Results: All values for McNemar and Paired t-test are significant at $p < 0.000001$. Cohen's $d$ is calculated relative to the unconstrained base $Z_0$.

| **Retention** | **English (EN)** | | **Persian (FA)** | |
|---|---|---|---|---|
| **Level ($Z_v$)** | **Mean Diff.** | **Cohen's $d$** | **Mean Diff.** | **Cohen's $d$** |
| $Z_{10}$ | 0.0799 | 0.3396 | **0.1353** | **0.4177** |
| $Z_{20}$ | 0.0578 | 0.2693 | **0.1023** | **0.3453** |
| $Z_{30}$ | 0.0457 | 0.2251 | **0.0791** | **0.2851** |
| $Z_{40}$ | 0.0356 | 0.1861 | **0.0642** | **0.2462** |
| $Z_{50}$ | 0.0299 | 0.1602 | **0.0540** | **0.2151** |
| $Z_{60}$ | 0.0246 | 0.1395 | **0.0487** | **0.1994** |
| $Z_{70}$ | 0.0191 | 0.1132 | **0.0419** | **0.1782** |
| $Z_{80}$ | 0.0164 | 0.0969 | **0.0366** | **0.1582** |
| $Z_{90}$ | 0.0131 | 0.0795 | **0.0316** | **0.1423** |

### 4.1. Key Findings on Language Sensitivity

The empirical results in Table 10 reveal a distinct disparity in how LLMs handle reasoning compression across languages:

1. **Higher Magnitude of Loss in Persian:** At every single retention level, the Mean Difference in probability scores for Persian is nearly double that of English. For instance, at $Z_{10}$, Persian suffers a 13.53% drop in sufficiency compared to only 7.99% in English.

2. **Effect Size Persistence:** In English, the impact of compression becomes negligible ($d < 0.2$) as early as $Z_{40}$. In contrast, Persian maintains a small but significant effect size ($d \geq 0.2$) up to $Z_{50}$, indicating that Persian models require a higher token density to sustain stable reasoning performance.

3. **Statistical Robustness:** The consistent $p < 10^{-6}$ across all levels confirms that the performance degradation following information pruning is a robust phenomenon rather than stochastic variance.

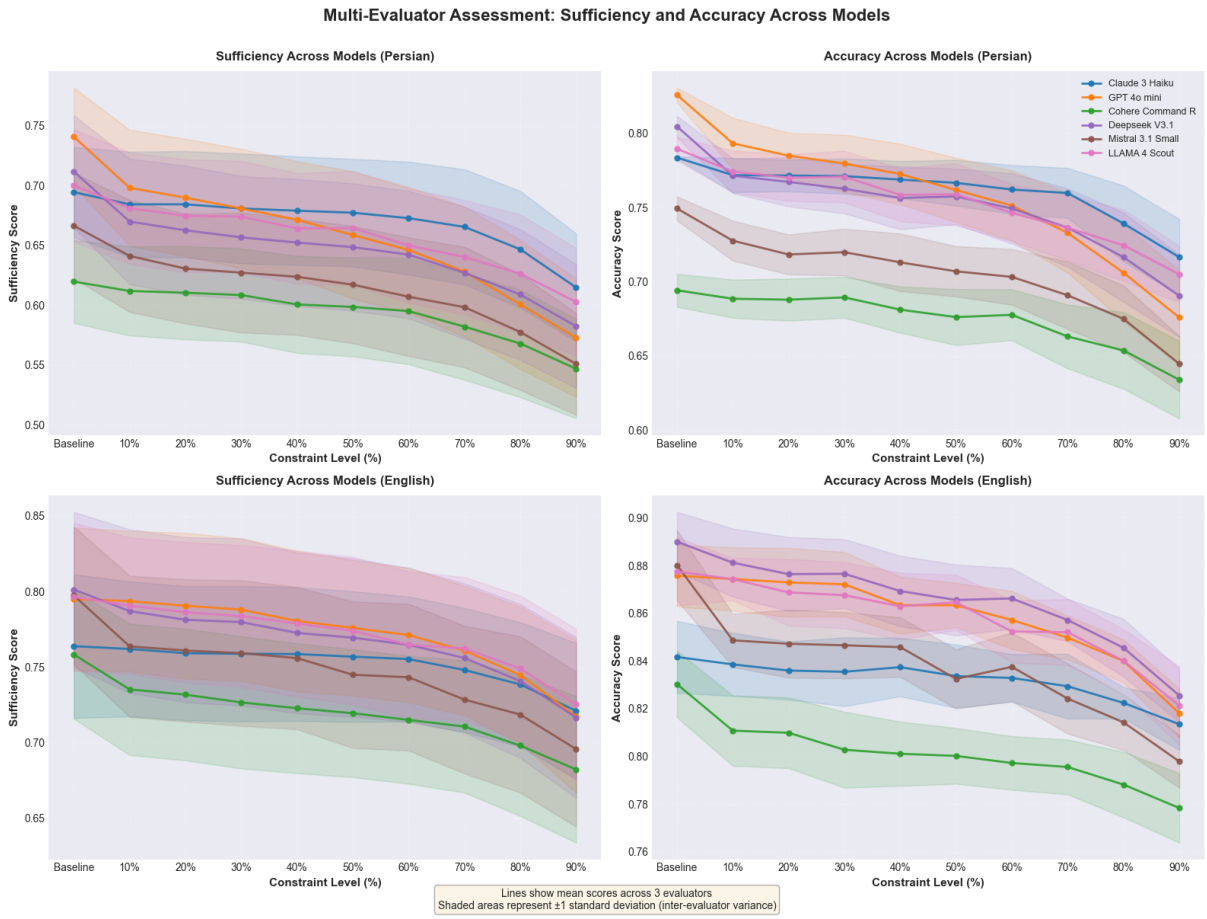


Figure 5: **Multi-evaluator assessment across constraint levels.** Performance of seven models on sufficiency (left) and accuracy (right) for Persian (top) and English (bottom). Solid lines show mean scores across three evaluators; shaded regions indicate ±1 standard deviation. Key findings: (1) GPT-4o mini and Claude 3 Haiku (orange, blue) maintain superior performance even at 90% constraint; (2) English models show greater resilience than Persian; (3) performance degradation accelerates beyond 60% constraint; (4) narrower bands in Persian indicate higher evaluator consensus.

## E. Multi-Evaluator Assessment Results

To ensure robust and unbiased quality assessment, we employed three independent evaluator models to score all generated explanations. This approach provides both aggregate performance metrics and inter-evaluator agreement measures, revealing where assessments are confident versus uncertain.

### 5.1. Performance Patterns

Figure 5 reveals several critical patterns. **Language differences:** English models demonstrate substantially better constraint resilience, maintaining sufficiency scores of 0.68–0.77 and accuracy scores of 0.78–0.82 even at 90% compression, compared to Persian's 0.51–0.62 and 0.63–0.68 respectively. The narrower confidence bands in Persian suggest higher inter-evaluator consensus, while wider bands in English (particularly at 40–60% constraints) indicate more nuanced quality assessments.

**Model hierarchy:** GPT-4o mini and Claude 3 Haiku consistently outperform other models across all conditions, with narrow confidence bands confirming evaluator consensus on their superiority. Cohere Command R consistently underperforms with the steepest degradation, particularly in sufficiency metrics.

**Critical threshold:** All models exhibit non-linear degradation, with performance decline accelerating significantly beyond the 60% constraint level. This suggests a critical information density point—below 40% of original length, essential information must be increasingly omitted, leading to rapid quality deterioration.

## 5.2. Inter-Evaluator Agreement

The confidence band analysis reveals important reliability patterns. High consensus emerges at baseline conditions and for extreme performers (best and worst models), indicated by narrow bands. Conversely, mid-range constraints (40–60%) produce wider bands, particularly for mid-tier models, suggesting an "uncertainty zone" where quality assessment becomes more subjective. Persian evaluations show approximately 10–15% narrower confidence bands on average, indicating more consistent quality criteria across evaluators.

## 5.3. Implications

For applications requiring constrained explanations, GPT-4o mini and Claude 3 Haiku are recommended, showing minimal quality loss even at 80% reduction. The 60% constraint level represents a critical threshold for deployment decisions. The multi-evaluator approach successfully quantifies not just performance levels but also confidence in those assessments, providing more actionable insights than single-evaluator metrics.