

# Can LLMs Faithfully Explain Themselves in Low-Resource Languages? A Case Study on Emotion Detection in Persian

Mobina Mehrazar<sup>1\*</sup>, Mohammad Amin Yousefi<sup>1\*</sup>,  
Parisa Abolfath Beygi<sup>2</sup>, Behnam Bahrak<sup>3</sup>

<sup>1</sup>School of Electrical and Computer Engineering College of Engineering, University of Tehran, Tehran, Iran

<sup>2</sup>Computer Science Department, University of British Columbia, Vancouver, Canada

<sup>3</sup>Tehran Institute for Advanced Studies, Khatam University, Tehran, Iran

m.amin.yousefi@ut.ac.ir, mobinamehrazar@ut.ac.ir, parisa.beygi@ubc.ca, b.bahrak@teias.institute

## Abstract

Large language models (LLMs) are increasingly used to generate self-explanations alongside their predictions, a practice that raises concerns about the faithfulness of these explanations, especially in low-resource languages. This study evaluates the faithfulness of LLM-generated explanations in the context of emotion classification in Persian, a low-resource language, by comparing the influential words identified by the model against those identified by human annotators. We assess faithfulness using confidence scores derived from token-level log-probabilities. Two prompting strategies, differing in the order of explanation and prediction (Predict-then-Explain and Explain-then-Predict), are tested for their impact on explanation faithfulness. Our results reveal that while LLMs achieve strong classification performance, their generated explanations often diverge from faithful reasoning, showing greater agreement with each other than with human judgments. These results highlight the limitations of current explanation methods and metrics, emphasizing the need for more robust approaches to ensure LLM reliability in multilingual and low-resource contexts.

**Keywords:** Large Language Model, Explanation Faithfulness, Model Calibration, Uncertainty Estimation, Low-Resource Languages, Emotion Detection

## 1. Introduction

The widespread adoption of large language models (LLMs), both open-source and proprietary, has increased the demand for interpretability in real-world applications. A key feature of many LLMs is their ability to generate natural language explanations for predictions, known as self-explanations.

Recent studies have introduced methods for generating and evaluating self-explanations in high-resource languages like English, primarily using auto-regressive LLMs (Huang et al., 2023; Madsen et al., 2024). The quality of these explanations depends on the model's grasp of textual context, linguistic nuances, and culturally embedded meanings (Liu et al., 2024).

An open question remains: in low-resource languages, where capturing linguistic nuances is more challenging, do self-explanations accurately represent the model's decision-making? With LLMs' growing capabilities in diverse linguistic contexts, addressing this is crucial.

To explore this, we evaluate the faithfulness of self-explanations in Persian, focusing on emotion detection. Using a Persian emotion-labeled corpus with formal and informal samples, we assess challenges in both language understanding and explanation generation.

Our analysis uses auto-regressive LLMs, which

identify keywords from the input as explanations for their predictions. We adopt two paradigms: Predict-then-Explain (P-E) and Explain-then-Predict (E-P) (Camburu et al., 2018), to explore the relationship between explanations and predictions. By isolating influential words, we assess the model's reasoning by measuring changes in predictions and confidence when these words are added or removed.

Previous work has often relied on self-reported confidence, where models express certainty through natural language explanations or specific words (Huang et al., 2023). However, empirical evidence shows that self-reported confidence estimates often diverge from calibrated confidence estimates, limiting their reliability for faithfulness evaluation (Zhou et al., 2024).

To overcome this limitation, we utilize token-level log-probabilities obtained through API access to compute more grounded confidence estimates (Xue et al., 2024). Throughout our analysis, we observe a consistent tendency toward overconfident predictions, which can distort faithfulness assessments. Accordingly, we apply temperature scaling (Guo et al., 2017), a post hoc calibration technique to improve the reliability of confidence estimates derived from token-level log-probabilities. These calibrated scores are then used to evaluate the faithfulness of self-explanations through a set of quantitative metrics.

While recent studies have explored the alignment between human reasoning and model-generated

\*These authors contributed equally to this work and are listed in alphabetical order.

explanations (Venkatesh et al., 2024), research is limited for low-resource languages. To address this, we investigate whether self-explanations in Persian align with human reasoning, a key step in evaluating LLM interpretability and reliability across diverse linguistic and cultural contexts. We designed a two-stage human annotation process that mirrors the output format of the models, enabling direct and fair comparisons with their explanations.

## 2. Related Work

### 2.1. Feature Attribution Methods

Recent work in explainable AI has focused on enhancing the interpretability of LLMs through feature attribution, which identifies input elements that most influence model outputs. Two main approaches have emerged. One involves analyzing the model’s sensitivity to small perturbations in input features, including gradient-based methods (Kindermans et al., 2016; Shrikumar et al., 2017), which trace relevance through internal model mechanisms. These methods are considered model-specific or analytical explanations. However, they typically require access to model weights, attention maps, or input gradients, assumptions that do not hold for closed-source LLMs, making them inapplicable in our setting.

Another approach to defining feature importance in LLMs is perturbation-based analysis, where input tokens, identified by model explanations, are altered to assess their impact on model predictions. Methods like LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) capture non-linear interactions by perturbing multiple tokens and assigning importance through linear regression (LIME) or Shapley values (SHAP). Recent adaptations for LLMs include TokenSHAP for Shapley-based token attribution (Horovicz and Goldshmidt, 2024) and studies on LIME’s explainability across model sizes (Heyen et al., 2024).

We adopt a model-compatible approach to feature attribution by prompting the LLM to generate self-explanations, identifying the most influential tokens for its prediction. Unlike gradient- or attention-based methods, this strategy directly elicits the model’s reasoning in natural language or extractive form. We evaluated the quality of these self-explanations by masking the identified tokens and measuring their impact on predictions. This work extends recent studies, as discussed in the following.

### 2.2. Self-Explanations via Input Token Attribution in LLMs

Randl et al. (2025) examines whether lightweight

LLMs (up to 8B parameters, like Gemma and Llama3-8B) can reliably generate self-explanations as influential words or phrases. The study, focusing on (P-E) paradigm, compares these explanations primarily with gradient-based methods. Although tasks like food hazard classification and sentiment analysis are explored, the lack of benchmarking on more widely-used models like GPT limits the generalizability to high-performing, real-world LLMs.

Huang et al. (2023) prompt ChatGPT to identify the top- $k$  most important tokens for binary sentiment classification and evaluate the faithfulness of these self-explanations using metrics like *comprehensiveness*, *sufficiency*, and *decision flip rate* (DFMIT). They compare self-explanations with traditional methods like LIME (Slack et al., 2020) and occlusion (Li et al., 2017), exploring both E-P and P-E paradigms. Notably, their evaluation relies on confidence scores generated by the model itself, a potentially unreliable method not based on internal probabilities.

Venkatesh et al. (2024) offers a complementary perspective by using eye-tracking to compare human, ML, and LLM-derived explanations in text classification. They extract top words from each source, LLM attributions, attention-based models, and human gaze saliency, to measure alignment. Unlike our method, they rely on implicit gaze behavior rather than explicitly asking humans to identify influential tokens, marking a key methodological distinction.

Madsen et al. (2024) uses a redaction-based consistency framework to assess explanation reliability, focusing on prediction flips. While it explores counterfactual explanations, it lacks models like GPT and human annotations, distinguishing it from our study.

### 2.3. Research Gap

To the best of our knowledge, there is no existing work that performs confidence estimation using the model’s log probabilities for evaluating self-explanations and subsequently calibrating these estimates to improve explanation quality. Furthermore, research on emotion detection in Persian, a resource-limited language, is lacking, which this study aims to address.

## 3. Methodology

This section outlines the experimental design used to evaluate the faithfulness of self-explanations generated by LLMs for emotion classification in Persian. We detail our interaction protocol with LLMs, the explanation extraction strategies employed, the approach to estimating model confidence and calibrating confidence, the structure of the human anno-

tation process, and the evaluation metrics used to assess explanation faithfulness and human agreement.

### 3.1. Interaction Protocol for Auto-regressive LLMs

Auto-regressive LLMs generate text sequentially, predicting one token at a time based on prior context. This left-to-right process enhances coherence and context awareness, making them well-suited for natural language generation. Built on transformer architectures, these models use self-attention to capture long-range dependencies.

In our experiments, we accessed each LLM through its API using a standardized prompting protocol with three distinct roles:

- **System:** specifies high-level instructions that govern the behavior of the model across all tasks.
- **User:** provides task-specific queries, prompts, or instructions.
- **Assistant:** generates responses, including classifications and explanations, based on the given instructions and queries.

This structured protocol ensures consistency in model behavior and output formatting throughout all stages of evaluation.

### 3.2. Explanation Extraction Approaches for the Classification Task

We use LLMs for emotion classification on a Persian-labeled corpus while generating self-explanations that identify influential words supporting each prediction. To examine how the order of explanation and prediction affects model behavior and explanation quality, we adopt two prompting paradigms from [Camburu et al. \(2018\)](#):

- **Explain-then-Predict (E-P):** The model first extracts the top- $k$  most influential words from the input text and outputs them as an explanation. In a subsequent turn of the same conversation, the model predicts the emotion label while having access to these words through the dialogue context.
- **Predict-then-Explain (P-E):** The model first predicts the emotion label from the input text and then generates an explanation by identifying the top- $k$  influential words supporting its decision.

Under both paradigms, the model assigns each input to one of six emotion categories: Sadness (0), Happiness (1), Anger (2), Surprise (3), Hatred (4),

and Fear (5), returning only the corresponding numeric label. Explanations are restricted to the top- $k$  influential words appearing in the original Persian input to ensure textual alignment.

In the E-P setting, the influential words are generated in an earlier turn and remain available through the chat history. During the classification step, the model receives the full input text together with this contextual information (see Table 12 in Appendix D). In contrast, in the P-E setting, the model first classifies the full input and then produces the influential words explaining its prediction (see Table 11 in Appendix D).

To evaluate the faithfulness of these explanations, we measure model predictions and confidence under three input formats:

- **Full-text:** the original input text.
- **Top- $k$  only:** the input contains only the extracted influential words.
- **Top- $k$  removed:** all occurrences of the influential words in the original input are replaced with a fixed placeholder token [removed]<sup>1</sup>.

Each condition is evaluated in a *fresh conversation* consisting only of system and user messages to avoid contamination from previous turns.

If the model maintains its prediction when only the influential words are provided, the explanation is considered *sufficient*. Conversely, substantial changes in prediction or confidence after removing these words indicate that they capture important evidence used in the decision process. These effects are quantified using the faithfulness metrics described in Section 3.5.

### 3.3. Confidence Estimation and Calibration

Evaluating the faithfulness of self-generated explanations requires reliable confidence estimates in the model's predictions. We use token-level log-probabilities of predicted labels, transforming them into normalized confidence scores (0 to 1).

To address the overconfidence of LLMs' probability estimates, we apply temperature scaling ([Guo et al., 2017](#)), a post-hoc calibration method. This introduces a scalar parameter, *Temperature*, optimized on a held-out calibration set by minimizing the *Expected Calibration Error* (ECE). The optimal *temperature* is then applied to the model's output probabilities during evaluation, improving the reliability of confidence estimates for faithfulness assessment.

<sup>1</sup>Persian equivalent used in the modified text: حذف شده.

The ECE is computed as:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (1)$$

where  $B_m$  denotes the set of predictions whose confidence scores fall into the  $m$ -th bin. To construct the bins, the confidence interval  $[0, 1]$  is partitioned into  $M$  equal-width sub-intervals

$$I_m = \left[ \frac{m-1}{M}, \frac{m}{M} \right), \quad m = 1, \dots, M,$$

and each prediction with confidence  $\hat{p}_i$  is assigned to bin  $B_m$  if  $\hat{p}_i \in I_m$ . Here,  $n$  is the total number of samples,  $\text{acc}(B_m)$  is the average accuracy of bin  $B_m$ , and  $\text{conf}(B_m)$  is the average confidence in that bin.

### 3.4. Human Annotation Process

To allow for a fair comparison with the model, we implemented a two-stage human annotation process that closely mirrors the structure of the LLM-generated outputs.

Twenty-five native Persian speakers participated as annotators, providing emotion classifications and influential words in a format consistent with the model’s output. Unlike the model, annotators selected the top- $k$  influential words and classified texts without a specific order, reflecting human decision-making. We did not require adherence to the E-P or P-E paradigms or collect explicit confidence scores, as these are not naturally expressed in human annotation.

In the first stage, each annotator received a distinct subset of full-text samples, classifying each into one of six emotion categories and identifying the top- $k$  influential words guiding their decision.

In the second stage, annotators classified two modified versions of previously unseen samples: (1) one with only the top- $k$  influential words, and (2) one with the top- $k$  words replaced by a placeholder token. Each version was classified into one of the six emotion categories.

To prevent familiarity bias and ensure decisions were based solely on the presented text, each annotator was assigned a unique set of samples in both stages. Additionally, all annotations underwent secondary verification, where they were cross-checked and validated by other annotators for consistency and correctness.

### 3.5. Evaluation Metrics

To evaluate the faithfulness of explanations generated by the underlying LLMs and human annotators, we employ a set of established evaluation metrics. In addition, we measure the pairwise agreement

between model- and human-generated explanations. The metrics used in this study are detailed below.

#### 3.5.1. Faithfulness Metrics

**Comprehensiveness** measures the impact of the most influential words on a model’s prediction by removing them and observing the change in confidence (DeYoung et al., 2020). The score is the difference in confidence before and after removal, with a higher score indicating a more faithful explanation.

**Sufficiency** measures whether the most influential words alone can drive the model’s prediction with the same confidence (DeYoung et al., 2020). It is calculated as the difference in confidence between the full-text input and the input containing only the influential words. A lower score indicates higher *sufficiency*, meaning the influential words alone are nearly sufficient for the prediction.

**Decision Flip Rate (Top- $k$  Only)** measures the model’s reliance on broader context by quantifying how often its prediction changes when only the top- $k$  influential words are retained. Noted as  $\text{DF}_{\text{TopKOnly}}$ , it averages the prediction flips across all samples. Higher values indicate that the selected words alone are insufficient to reproduce the original decision, suggesting a greater need for contextual information.

**Decision Flip Rate (Top- $k$  Removed)** quantifies how often a model’s prediction changes when its most influential words are replaced with placeholders (Chrysostomou and Aletras, 2021). Noted as  $\text{DF}_{\text{TopKRemoved}}$ , it calculates the average rate of prediction flips across the dataset, reflecting the model’s sensitivity to key input words. Higher values indicate greater reliance on these words, highlighting their critical role in decision-making.

#### 3.5.2. Agreement Metrics

**Feature Agreement** measures the overlap between the top- $k$  most influential features identified by two different explanation methods. It is calculated as the fraction of common features between the two sets of top- $k$  features. Higher values indicate greater alignment in what both sources consider important, suggesting stronger agreement in explanatory focus (Krishna et al., 2024).

**Intersection-over-Union (IoU)** measures the similarity between two sets of top- $k$  influential features by comparing their intersection relative to their union. It is defined as

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where  $A$  and  $B$  are the feature sets under comparison. Values range from 0 (no overlap) to 1

Model	Predict-then-Explain (P-E)				Explain-then-Predict (E-P)			
	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy
<b>GPT-4o</b>	<b>81</b>	<b>78</b>	<b>77</b>	<b>77</b>	<b>78</b>	<b>75</b>	<b>74</b>	<b>75</b>
<b>GPT-4-turbo</b>	75	72	70	72	74	68	67	68
<b>Llama3.3-70B-instruct</b>	74	65	61	65	79	62	59	62
<b>DeepSeek-V3</b>	58	65	60	65	71	62	59	62
<b>Human Annotators</b>	69	66	67	66	69	66	67	66

Table 1: Comparison of classification performance in P-E and E-P settings. All metrics are reported as percentages.

(perfect overlap). Higher values indicate stronger agreement between the two explanation methods in terms of the features they identify as important.

## 4. Experiments and Results

### 4.1. Experimental Setup

Following our methodology, we selected a set of LLMs that offer access to token-level log-probabilities via their official APIs. These models vary in architecture and scale, with parameter counts ranging from approximately 70 billion to over one trillion. Specifically, our experiments used Llama3.3-70B-instruct (Grattafiori et al., 2024) (70B parameters), DeepSeek-V3 (et al., 2025) (671B parameters), GPT-4-turbo (et al., 2024a), and GPT-4o (et al., 2024b), both estimated to have over one trillion parameters.

We conducted our experiments on a subset of the ARMANEMO dataset (Mirzaee et al., 2025), a manually annotated Persian corpus for emotion detection in both formal and colloquial text. ARMANEMO comprises over 7,000 sentences labeled with six emotion categories and an additional OTHER class for out-of-scope emotions. To ensure focused evaluation, we excluded Other-labeled samples and curated a balanced evaluation set of 300 instances (about 50 per emotion). An additional disjoint set of 210 samples (35 per emotion) was used for temperature scaling to calibrate model confidence.

To select an appropriate value for  $k$ , we conducted a small-scale empirical analysis on a subset of the data. While dynamic values based on text length were initially considered, they often underperformed: many short texts contained multiple influential words, and low  $k$  values failed to capture them all. To address this, we chose a fixed  $k = 5$  for all samples. Although this may introduce less informative words in shorter texts, it ensures emotionally salient words are not missed, which is crucial for generating faithful and comprehensive explanations.

### 4.2. Emotion Classification Performance

Table 1 reports macro-level classification performance across the E-P and P-E prompting paradigms. Consistently, P-E outperforms E-P across all models, suggesting that predicting before explaining leads to better results, echoing findings by Huang et al. and Camburu et al..

As shown, ChatGPT-family models outperform even human annotators, particularly in terms of accuracy and F1-score. In contrast, Llama3.3-70B-instruct and DeepSeek-V3 underperform relative to humans with a slight difference, highlighting current limitations in their ability to interpret Persian emotional content comparing to ChatGPT-family models. The comparatively lower performance of human annotators does not necessarily indicate inferior emotional interpretation but rather highlights a fundamental characteristic of the evaluation benchmark. Many texts in the dataset express a mix of emotions, where the 'gold' label represents a majority vote. An LLM, as a probabilistic model, is optimized to predict the most statistically likely outcome, i.e. the majority label. In contrast, a human annotator might legitimately identify a valid but secondary emotion and be marked as 'incorrect' by a benchmark that only accepts a single answer. Thus, the models' superior scores may reflect their alignment with the benchmark's majoritarian nature more than a 'truer' understanding of the emotional content.

Table 2 provides a detailed comparison of model performance by emotion category under the E-P and P-E settings. GPT-family models show relatively stable results across both paradigms, indicating robustness in handling subtle or context-dependent emotions. Generally, E-P yields higher F1-scores for Happiness, Fear, and Sadness, while P-E performs better for Anger and Hatred. The Surprise category poses consistent challenges across models, likely due to its nuanced and context-sensitive expression in Persian, often involving tone, sarcasm, or ambiguity. Misclassifications of Surprise frequently map to Anger or Happiness. Notably, DeepSeek-V3 fails to correctly classify any samples from the Surprise category in the E-

Model	Metric	Sadness		Happiness		Anger		Surprise		Hatred		Fear	
		P-E	E-P	P-E	E-P	P-E	E-P	P-E	E-P	P-E	E-P	P-E	E-P
GPT-4o	Precision	79	70	94	96	58	55	92	95	90	70	74	81
	Recall	63	71	90	96	92	76	46	40	75	78	100	88
	F1-Score	70	70	92	96	71	64	61	56	82	74	77	84
GPT-4-turbo	Precision	81	74	90	89	51	42	81	94	79	55	69	87
	Recall	25	54	86	96	82	49	60	34	86	92	94	85
	F1-Score	38	62	88	92	63	45	69	50	82	69	80	86
Llama3.3-70B-instruct	Precision	68	79	45	100	65	89	69	77	37	100	100	94
	Recall	69	96	90	04	63	71	67	94	100	10	35	65
	F1-Score	69	86	60	08	64	79	68	85	54	18	52	77
DeepSeek-V3	Precision	66	75	79	83	43	35	00	100	71	48	91	87
	Recall	73	58	96	96	78	43	00	08	82	88	62	81
	F1-Score	69	65	86	89	55	39	00	15	76	62	74	84
Human Annotators	Precision	62		95		50		71		55		82	
	Recall	63		82		67		54		67		65	
	F1-Score	63		88		57		61		60		72	

Table 2: Comparison of classification Performance across P-E and E-P paradigms for each emotion separately on full-text samples. All metrics are reported as percentages.

P setup, highlighting the difficulty of detecting this emotion class. Additionally, a high degree of confusion between Anger and Hatred suggests substantial semantic and affective overlap, complicating their disambiguation.

### 4.3. Process for Collecting Explanations

To collect self-explanations, both language models and human annotators were asked to identify the top 5 most influential words contributing to their emotion classification. All selected words had to exactly match those in the original input.

### 4.4. Improving Confidence Estimation

To improve the reliability of confidence estimates for explanation faithfulness, we applied temperature scaling independently for each model and prompting paradigm. A grid search over the range [0.1, 21.0] with a 0.1 step size identified the optimal *temperature* on a calibration set, which was then applied to the evaluation set.

Table 3 shows that temperature scaling significantly reduces ECE across all models and prompting paradigms. These improvements generalize well to the evaluation set, indicating that the calibrated temperatures effectively correct the models' overconfident predictions.

Model	Dataset & Scaling	ECE	
		P-E	E-P
GPT-4o	Pre-Scale Calib. Set	23.26	25.04
	Post-Scale Calib. Set	6.20	6.68
	Post-Scale Eval. Set	4.83	6.64
GPT-4-turbo	Pre-Scale Calib. Set	36.96	37.22
	Post-Scale Calib. Set	6.90	3.57
	Post-Scale Eval. Set	4.99	11.52
Llama3.3-70B-instruct	Pre-Scale Calib. Set	43.19	47.31
	Post-Scale Calib. Set	2.36	3.57
	Post-Scale Eval. Set	1.40	4.15
DeepSeek-V3	Pre-Scale Calib. Set	39.26	37.40
	Post-Scale Calib. Set	7.73	5.78
	Post-Scale Eval. Set	5.96	5.57

Table 3: Temperature optimization and calibration results across models, reported as percentages.

### 4.5. Faithfulness Evaluation

We assess how explanation sequencing affects the faithfulness of self-generated explanations by comparing the E-P and P-E prompting paradigms across different LLMs.

Faithfulness is evaluated under both pre- and

Model	Comp ( $\uparrow$ )	Suff ( $\downarrow$ )	DF <sub>TopKRemoved</sub> ( $\uparrow$ )	DF <sub>TopKOnly</sub> ( $\downarrow$ )
<b>Predict-then-Explain (P-E)</b>				
<b>GPT-4o</b>	<b>27.42</b>	8.08	38.00	<b>22.00</b>
<b>GPT-4-turbo</b>	17.76	<b>-1.57</b>	<b>53.33</b>	24.33
<b>Llama3.3-70B-instruct</b>	7.96	7.84	30.33	29.33
<b>DeepSeek-V3</b>	18.80	12.74	39.67	29.00
<b>Explain-then-Predict (E-P)</b>				
<b>GPT-4o</b>	22.82	24.76	31.67	36.00
<b>GPT-4-turbo</b>	24.36	24.56	43.00	37.67
<b>Llama3.3-70B-instruct</b>	12.22	<b>18.59</b>	27.00	42.00
<b>DeepSeek-V3</b>	<b>28.77</b>	22.08	<b>48.33</b>	<b>30.00</b>
<b>Human Annotators</b>	-	-	53.33	39.00

Table 4: Faithfulness comparison of underlying LLMs in P-E and E-P settings, after confidence calibration. All metrics are reported as percentages.

post-calibration settings. As discussed in Section 4.4, calibration improves the reliability of confidence scores, notably enhancing *Sufficiency* in the P-E setting. This effect is primarily due to temperature scaling reducing overall confidence values, with higher-confidence predictions being scaled down more aggressively.

Since predictions based on the full-text input typically have higher confidence than those based on either the top- $k$  words only or the top- $k$  words removed, calibration narrows the confidence gap between these modes. As a result, *Sufficiency*, which measures how well the top- $k$  words alone preserve the model’s confidence, tends to increase. However, the same compression in confidence scores reduces the gap between the full-text input and the top- $k$  words removed input, leading to lower *Comprehensiveness*.

As shown in Table 4, most models exhibit greater explanation faithfulness under the P-E setup. This aligns with prior findings on English-language datasets (Huang et al., 2023), suggesting that generating explanations after predictions yields more faithful results, even for Persian emotion classification. DeepSeek-V3 is a notable exception, performing better when influential words are identified before prediction. However, Llama3.3-70B-instruct consistently shows weaker faithfulness across both paradigms, with a notable decline in *comprehensiveness* after calibration, possibly due to its significantly smaller size compared to the other models.

In the P-E setup, While no single model outperforms across all metrics, clear patterns emerge: GPT-4o leads in *comprehensiveness* and DF<sub>TopKOnly</sub>, whereas GPT-4-turbo excels in *sufficiency* and DF<sub>TopKRemoved</sub>. These results suggest that both models are more effective in identifying the most influential words in their predictions. But

Unexpectedly, in the E-P setting, DeepSeek-V3 outperforms other models in most metrics.

Interestingly, and contrary to expectations, GPT-4-turbo produces more confident predictions when given only the identified influential words, without additional context.

Compared to the models, human annotators perform well on DF<sub>TopKRemoved</sub>, indicating that their emotion predictions rely heavily on the identified influential words, and removing these often changes their decisions.

#### 4.6. Agreement Evaluation between LLM and Human Reasoning

Figure 1 shows the agreement between model-generated and human-annotated top- $k$  words, using Feature Agreement and Intersection over Union (IoU) as metrics. These scores are computed under both E-P and P-E prompting paradigms. To ensure fair comparison, agreement is measured only on instances where both agents assigned the same label to the input text.

Agreement scores are consistently higher in the P-E setting, indicating that post-prediction explanations align better with other models and human annotators. This highlights the impact of prompting order, motivating our exclusive focus on the P-E setting in subsequent analyses.

Model-to-model agreement exceeds 50%, reflecting moderate consistency in identifying top- $k$  influential words. GPT-4o and DeepSeek-V3 exhibit the highest mutual agreement, suggesting similar interpretive strategies. In contrast, model-human agreement is consistently lower, highlighting fundamental differences in how humans and models prioritize influential words.

The consistently higher model-to-model agree-

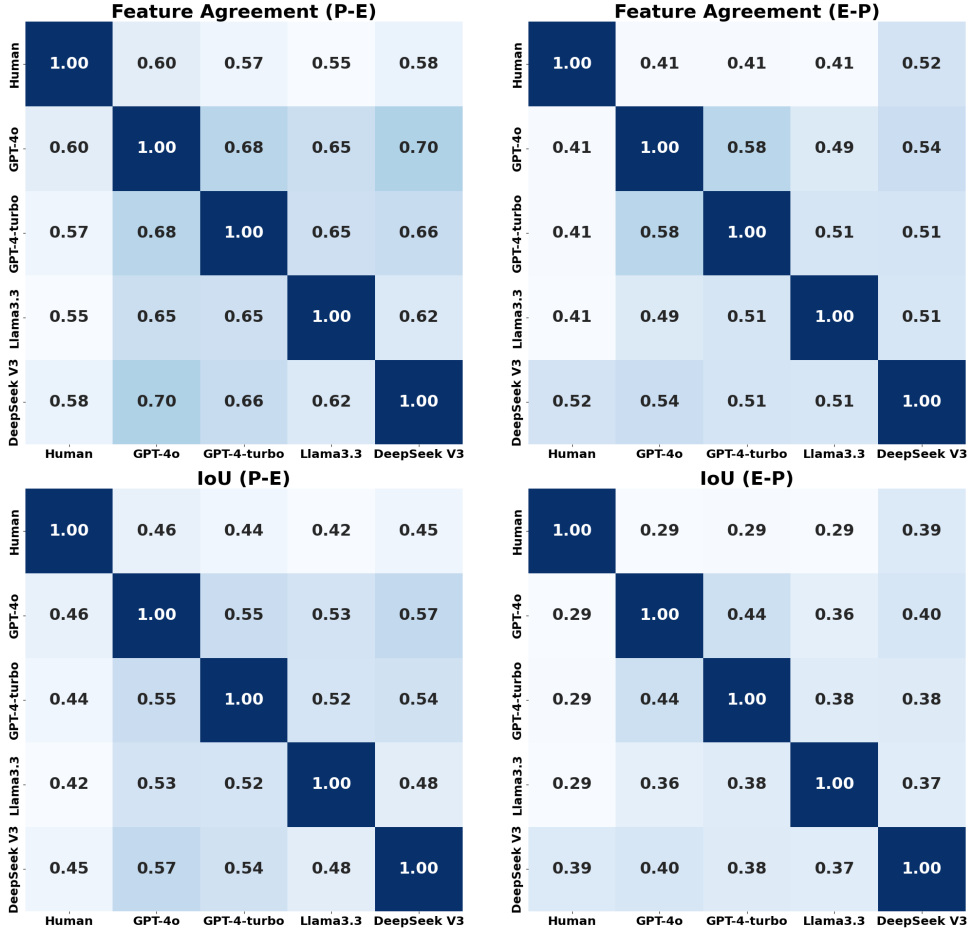


Figure 1: Comparison of explanation agreement metrics between human annotators and model explanations under both P-E and E-P prompting paradigms.

ment compared to model-human agreement suggests that LLMs may be converging on similar reasoning strategies, even when trained independently. This likely reflects their reliance on shared statistical regularities or keyword-based heuristics present in the training data, which makes their explanations mutually consistent. Humans, by contrast, draw on semantic understanding, pragmatic cues, and cultural context when identifying influential words, leading to explanations that often diverge from model-generated rationales. This divergence underscores a central challenge in explainability: achieving faithfulness requires more than internal model alignment; it demands bridging the gap between statistical shortcuts and the richer, context-driven reasoning used by humans. Notably, this finding reinforces our earlier hypothesis in Sec. 4.2, where we argued that LLMs’ superior benchmark performance reflects alignment with majoritarian statistical patterns rather than deeper semantic understanding. Taken together, these results highlight a systematic difference in how humans and LLMs reason about text. The challenge is especially pronounced in low-resource settings, where limited training data may aggravate reliance on shallow cor-

relations, further widening the gap between model and human explanations.

#### 4.7. Qualitative Analysis

While quantitative metrics offer useful benchmarks, they cannot fully capture why model explanations diverge from human reasoning, even when both yield the same label. Our qualitative examination of representative Persian samples reveals a consistent trend: humans emphasize affective or colloquial cues, whereas models attend to semantically central but emotionally neutral words. This reflects a broader limitation of large language models, their reliance on lexical salience over socio-emotional understanding. Such divergence is particularly pronounced in low-resource, colloquial-rich languages like Persian, underscoring the need for models that better integrate pragmatic and cultural cues.

### 5. Conclusion

We assessed the faithfulness of self-generated explanations from several auto-regressive LLMs using

token-level log-probability analysis. Models were prompted to identify their top- $k$  influential words in emotion classification tasks on a Persian dataset, comparing the prompting paradigms E-P and P-E. Faithfulness metrics and agreement with human annotations were used for evaluation.

Temperature scaling improved model calibration, enhancing the alignment of model confidence with prediction reliability. GPT-family models outperformed native Persian annotators, showcasing their strength in low-resource languages, though no model achieved high faithfulness metrics. This suggests either the models' limitations in producing faithful explanations or the inadequacy of the chosen explanation strategy and evaluation metrics.

Models showed higher agreement with each other than with human annotations, indicating a divergence in reasoning. The P-E paradigm yielded more accurate classifications, faithful explanations, and closer alignment with human annotations, suggesting that post-hoc rationalization improves explanation quality.

These findings highlight the importance of model architecture and prompting strategy in generating faithful self-explanations. Future work should explore alternative explanation strategies and refine faithfulness metrics, as well as expand research to diverse datasets, languages, and tasks for more robust evaluations.

## 6. Limitations

Our study has several limitations. First, we could not evaluate locally hosted models due to the high computational and memory demands of large-scale white-box models. Additionally, few models offer token-level log-probabilities through official APIs, limiting comparative analysis.

Second, our prompting paradigm introduces limitations. Although we used consistent prompts across models, LLMs' sensitivity to phrasing means different formulations could yield different predictions or explanations. Their inherent stochasticity may also lead to slight variations across runs, even under identical settings. While we carefully designed the prompts, we cannot guarantee perfect alignment with model interpretation.

Third, our formulation of emotion classification inherits the dataset's single-label setup, which may be an overly simple operationalization of emotion. Many texts plausibly express mixed or overlapping affective states, and it is not clear that humans naturally judge such inputs according to a single discrete label (or a one-dimensional notion of emotion), which can affect both human-model comparisons and agreement measurements.

Fourth, our explanation format is intentionally constrained to top- $k$  influential words, which may

under-specify the rationale and make comparisons with human explanations less explicit. In particular, the observed tendency of models to agree more with each other than with humans may partly reflect shared training data characteristics and similar learning paradigms across contemporary LLMs, whereas humans rely on substantially different experiential data and less well-characterized learning processes, an asymmetry our current setup does not isolate.

Finally, the metrics employed are proxies and may not fully capture alignment with the models' internal reasoning. Therefore, our findings do not suggest that LLMs are incapable of faithful self-explanation, but rather that, under the specific strategies and metrics used in this study, they often fall short.

## 7. Bibliographical References

- Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-SNLI: Natural Language Inference with Natural Language Explanations](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.
- George Chrysostomou and Nikolaos Aletras. 2021. [Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 477–488, Online. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.
- DeepSeek-AI et al. 2025. [DeepSeek-V3 Technical Report](#).
- OpenAI et al. 2024a. [GPT-4 Technical Report](#).
- OpenAI et al. 2024b. [GPT-4o System Card](#).
- Aaron Grattafiori, Abhimanyu Dubey, Hugo Touvron, Angela Fan, Gabriel Synnaeve, Sergey Edunov, Armen Aghajanyan, Thomas Scialom, Roberta Raileanu, Soumith Chintala, and Meta AI. 2024. [The Llama 3 Herd of Models](#).

- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On Calibration of Modern Neural Networks](#). In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.
- Henning Heyen, Amy Widdicombe, Noah Yamamoto Siegel, Philip Colin Treleaven, and Maria Perez-Ortiz. 2024. [The Effect of Model Size on LLM post-hoc Explainability via LIME](#). In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- Miriam Horovicz and Roni Goldshmidt. 2024. [TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation](#). In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 1–8, Miami, FL, USA. 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](#).
- Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. [Investigating the Influence of Noise and Distractors on the Interpretation of Neural Networks](#).
- Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. [The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective](#). *Transactions on Machine Learning Research*.
- Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. [Understanding Neural Networks through Representation Erasure](#).
- Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. [Are Multilingual LLMs Culturally-diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. 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*, volume 30. Curran Associates, Inc.
- 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*, pages 295–337, Bangkok, Thailand. Association for Computational Linguistics.
- Hossein Mirzaee, Javad Peymanfar, Hamid Habibzadeh Moshtaghin, and Hossein Zeinali. 2025. [ArmanEmo: a Persian Dataset for Text-based Emotion Detection](#). *Language Resources and Evaluation*.
- Korbinian Randl, John Pavlopoulos, Aron Henriksen, and Tony Lindgren. 2025. [Evaluating the Reliability of Self-explanations in Large Language Models](#). In *Discovery Science*, pages 36–51, Cham. Springer Nature Switzerland.
- 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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. [Learning Important Features Through Propagating Activation Differences](#). In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.
- Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. [Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods](#). In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA. Association for Computing Machinery.
- Jeevithashree Divya Venkatesh, Aparajita Jaiswal, and Gaurav Nanda. 2024. [Comparing Human Text Classification Performance and Explainability with Large Language and Machine Learning Models Using Eye-tracking](#). *Scientific Reports*, 14:14295.
- Boyang Xue, Hongru Wang, Rui Wang, Sheng Wang, Zezhong Wang, Yiming Du, Bin Liang, and Kam-Fai Wong. 2024. [A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models](#).
- Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. [Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Model	Input Format	Average Reduction	
		P-E	E-P
GPT-4o	Full-text	16.60	12.17
	Top- <i>k</i> only	8.77	7.07
	Top- <i>k</i> removed	12.32	9.85
GPT-4-turbo	Full-text	39.03	20.65
	Top- <i>k</i> only	17.85	10.95
	Top- <i>k</i> removed	9.67	5.03
Llama3.3 70B-instruct	Full-text	36.82	35.22
	Top- <i>k</i> only	18.71	14.99
	Top- <i>k</i> removed	17.71	22.71
DeepSeek-V3	Full-text	33.16	25.38
	Top- <i>k</i> only	21.39	24.89
	Top- <i>k</i> removed	18.48	14.34

Table 5: Average model confidence score reduction of before and after temperature scaling on the evaluation set. Results are reported across different input formats (full-text, top-*k* only, and top-*k* removed) and prompting paradigms (P-E and E-P). Larger reductions reflect stronger correction of overconfident predictions. The results are reported as percentages.

### A. Human Annotation Protocol and Participant Details

Human annotations were collected from a group of 25 native Persian speakers, all undergraduate students in computer science and computer engineering. Participants included both male and female students and were recruited from an academic environment on a voluntary basis. All participants were informed about the purpose of the study and how their annotations would be used.

All annotations were gathered using Google Forms. In the first stage, annotators were assigned unique subsets of full-text samples. For each instance, they classified the text into one of six emotion categories and selected the top-*k* influential words guiding their decision, following translated versions of the same instructions provided to the LLMs.

In the second stage, annotators were presented with two modified versions of previously unseen samples: (1) a sparse-input version containing only the top-*k* words, and (2) a masked-input version in which the top-*k* words were replaced by a placeholder token. As in the first stage, they classified each input into one of the six emotion categories.

To maintain annotation quality and prevent fa-

miliarity bias, no annotator saw the same instance across both stages. All annotations were independently verified through cross-checking by fellow annotators to ensure consistency and correctness. All participants consented to contribute voluntarily, and their responses were anonymized.

### B. Calibration Impact on Confidence Scores

Table 5 summarizes the average reduction in confidence scores across input types and prompting paradigms. Reductions were most pronounced for full-text inputs, producing the largest confidence decreases.

### C. Faithfulness Evaluation Before Calibration

Table 6 presents the self-explanation faithfulness metrics for all models under both prompting paradigms, prior to confidence calibration. These results provide a baseline for understanding the raw behavior of each LLM before post-processing adjustments to confidence scores.

### D. Prompt Templates for API-Based LLM Interactions

This section outlines the standardized prompt templates used to interact with LLMs during evaluation. All interactions followed a structured role-based format involving `system`, `user`, and `assistant` messages.

Table 7 shows the prompt used for classifying the full input text. Table 8 illustrates how we extracted the top-*k* influential words (self-explanations), without requesting an emotion label. Table 9 presents the prompt used to classify based solely on the extracted influential words. Table 10 contains the prompt used for classification after masking these words in the original input.

Table 11 and 12 consolidates the full interaction flow under the Predict-then-Explain (P-E) and Explain-then-Predict (E-P) paradigm for a single evaluation sample, including the Persian input text, Chat GPT-4o model outputs, and word masking stages.

Note that in rare instances (fewer than 1%) where the model output did not conform to the expected format (e.g., producing additional text), these samples were discarded to ensure the integrity and consistency of our reported metrics.

Model	Comp ( $\uparrow$ )	Suff ( $\downarrow$ )	DF <sub>TopKRemoved</sub> ( $\uparrow$ )	DF <sub>TopKOnly</sub> ( $\downarrow$ )
<b>Predict-then-Explain (P-E)</b>				
GPT-4o	31.7	<b>15.91</b>	38.00	<b>22.00</b>
GPT-4-turbo	<b>47.11</b>	19.60	<b>53.33</b>	24.33
Llama3.3-70B-instruct	27.08	25.95	30.33	29.33
DeepSeek-V3	33.48	24.52	39.67	29.00
<b>Explain-then-Predict (E-P)</b>				
GPT-4o	25.13	29.86	31.67	36.00
GPT-4-turbo	<b>39.98</b>	34.26	43.00	37.67
Llama3.3-70B-instruct	24.73	38.81	27.00	42.00
DeepSeek-V3	39.82	<b>22.57</b>	<b>48.33</b>	<b>30.00</b>
<b>Human Annotators</b>	-	-	53.33	39.00

Table 6: Faithfulness comparison of underlying LLMs in P-E and E-P settings, before confidence calibration. All metrics are reported as percentages.

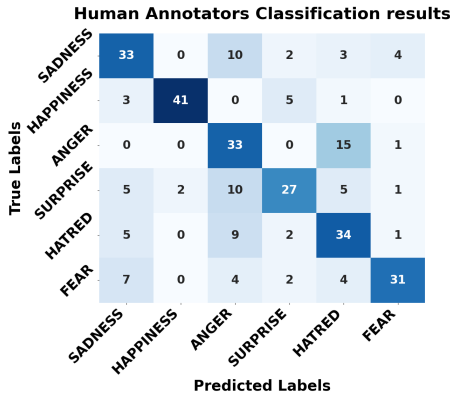


Figure 2: Confusion matrices of Human Annotator performance.

## E. Emotion Prediction Confusion Matrices

Figure 3 displays the confusion matrices for all evaluated models under both P-E and E-P prompting conditions. Figure 2 shows the corresponding confusion matrix for human annotators.

## F. Reliability Diagrams

To visually assess the impact of temperature scaling on confidence calibration, we present reliability diagrams for both the calibration and evaluation sets. For visualization, we used the implementation provided by Guo et al. (2017), available at their official GitHub repository<sup>2</sup> with 20 bins for all models and prompting paradigms.

Prior to calibration, the models exhibit significant

<sup>2</sup><https://github.com/hollance/reliability-diagrams?tab=readme-ov-file>

overconfidence. Most predictions have confidence scores close to 1, while actual accuracy lags behind, as shown by the noticeable gap between the red confidence bars and the black accuracy bars in the diagrams. This misalignment results in high *Expected Calibration Error* (ECE) values.

After applying temperature scaling, this discrepancy is substantially reduced. The confidence scores become more evenly distributed across the [0, 1] interval, and the gap between predicted confidence and actual accuracy narrows considerably. This leads to a marked reduction in ECE and a better-calibrated model, where confidence scores are more trustworthy indicators of predictive accuracy.

This improvement generalizes to the evaluation set as well. Using the *temperature* parameters optimized for each model on the calibration set, we observe a similar alignment between the confidence of each model and accuracy in evaluation, confirming the effectiveness of temperature scaling across datasets.

## G. Qualitative Examples of Human-Model Divergence

As shown in Table 6, GPT-family models (GPT-4o and GPT-4-turbo) consistently outperform Llama3.3-70B-instruct and DeepSeek-V3 in faithfulness metrics. For this reason, the qualitative examples presented here focus exclusively on GPT models in comparison with human annotations. This enables us to better isolate the key differences between state-of-the-art model explanations and human reasoning.

Across examples, we observe that models often prioritize topical or semantically central terms, while

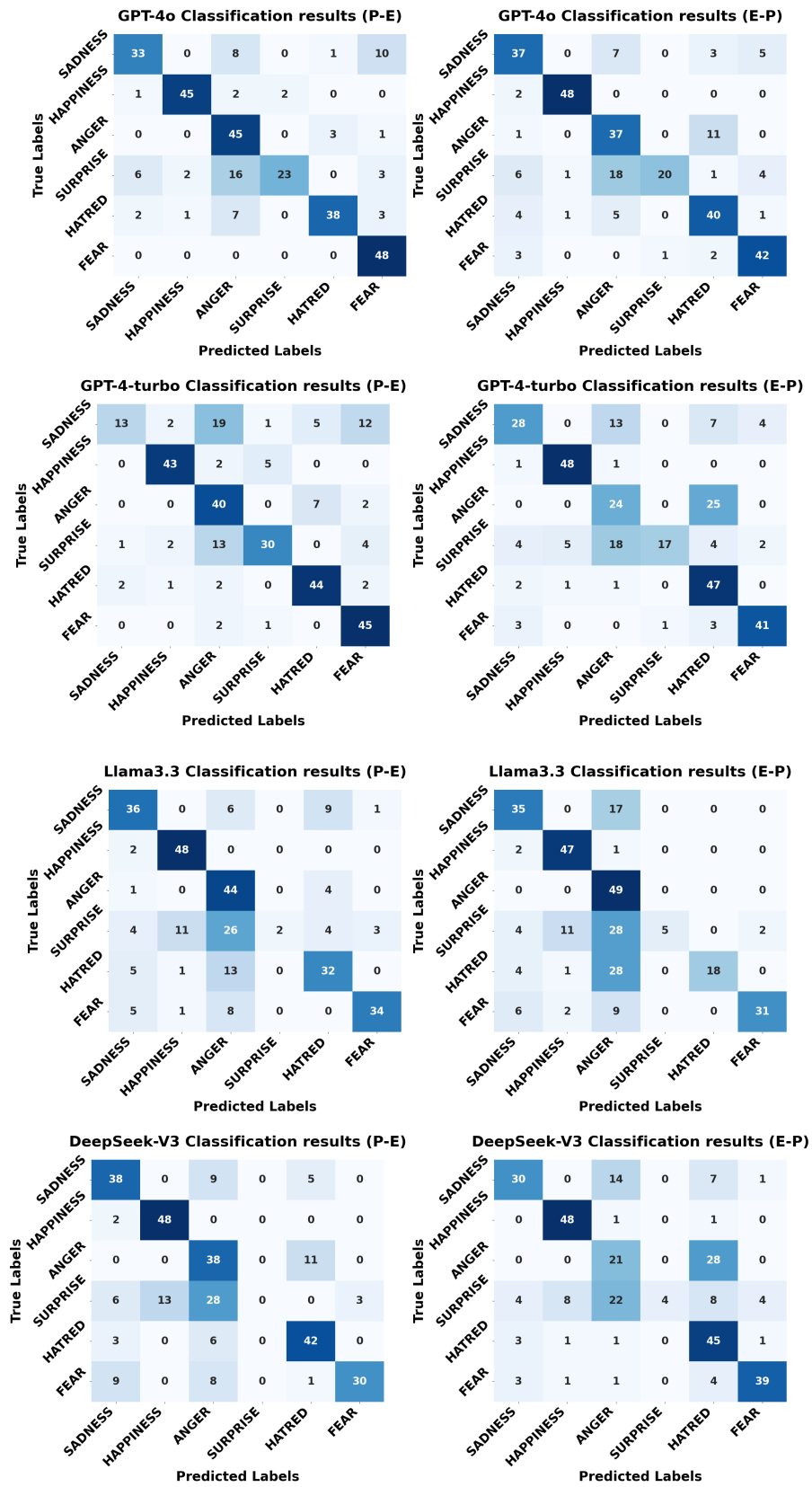


Figure 3: Confusion matrices of different models in both P-E and E-P paradigms

Role	Content
system	You are an emotion classifier. You must classify the emotion and output the top influential words in CSV format. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable. For top influential words, you can only output Persian words in the text.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class, output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character. Text:

Table 7: The prompt used for the LLM for emotion classification of the original text

Role	Content
system	You are an emotion classifier. You must classify the emotion and output the top influential words in CSV format. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable. For top influential words, you can only output Persian words in the text.
user	List the top K most influential words that contributed to this classification in CSV format (in a single line). I don't want you to classify in this stage. Only give me top K words. Make sure to provide only K Persian words which they exist in the original text and don't output any other token. Here is an example output: word1,word2,word3,word4,word5 Text:

Table 8: The prompt used for the LLM to get the most influential words from the original text as the self-explanation

human annotators highlight colloquial, affective, and contextually charged words. This divergence helps explain why model explanations, though computationally faithful, may still fail to align with human reasoning.

Table 13 presents representative samples comparing GPT-4-turbo and human annotations under both Predict-then-Explain (P-E) and Explain-then-Predict (E-P) settings. We observe that GPT-4-turbo frequently highlights structurally or thematically relevant words (e.g., کتابخون : "well-read", مامانت : "your mom"), whereas human annotators consistently point to affective slang or evaluative markers (e.g., خر : "tupid", گند : "mess") as decisive signals of emotion. Even in positive cases, such as texts labeled Happiness, humans emphasize intensifiers like بیش باد ("may it increase") that convey strong cultural and emotional resonance, while the model focuses on neutral topical references such as named entities.

Table 14 provides analogous examples for GPT-4o. Here we see a similar trend: the model tends to select broad or structural indicators of sentiment (e.g., مشکل : "problem", فساد : "corruption"), whereas humans privilege direct insults, sarcasm, and culturally salient cues (e.g., خر : "stupid", سگ : "dog"). In surprise or fear cases, GPT-4o often highlights neutral topical markers like نمودار ("chart") or انسانیت ("humanity"), while humans select emotionally charged expressions such as برگام ریخت ("I was shocked") or هیولا ("monster").

Taken together, these qualitative cases illustrate

that even the most underlying faithful GPT-family models remain misaligned with human reasoning when interpreting emotional language. The gap arises primarily from underweighting colloquial, affective, and culturally grounded expressions, features that are especially critical in low-resource and slang-rich languages like Persian.

Role	Content
system	You are an emotion classifier. You are provided with some influential words that have been extracted from the text. You must classify the emotion based only on these words. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character. Text:

Table 9: The prompt used for the LLM for emotion classification of the top-k most influential words alone

Role	Content
system	You are an emotion classifier. In the text, some influential words have been replaced with the placeholder [removed]. You must classify the emotion based on the text, considering these [removed] words as part of the context. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character. Text:

Table 10: The prompt used for the LLM for emotion classification after replacing the top-k most influential words in the original text with placeholders

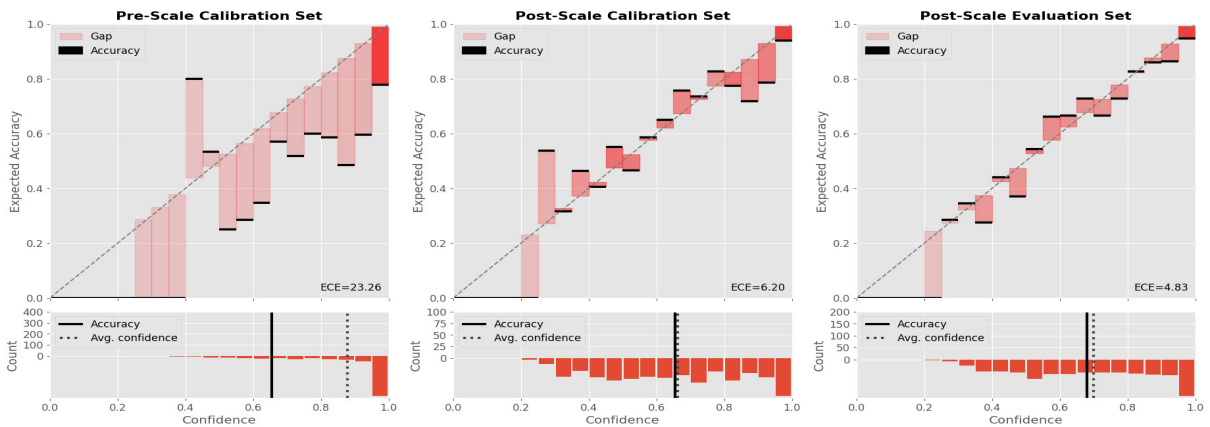


Figure 4: Reliability diagrams of the Chat GPT-4o model under P-E paradigm.

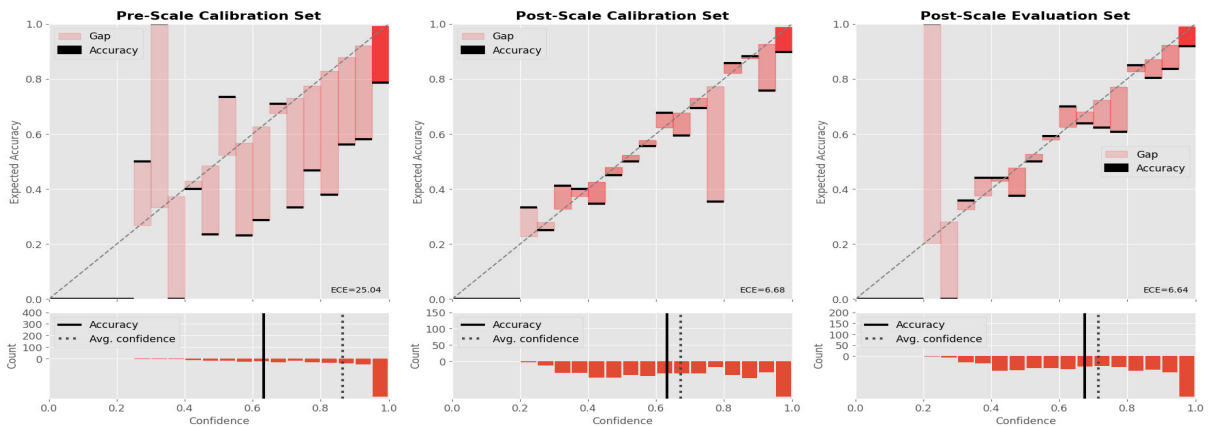


Figure 5: Reliability diagrams of the Chat GPT-4o model under E-P paradigm.

Role	Content
system	You are an emotion classifier. You must classify the emotion and output the top influential words in CSV format. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable. For top influential words, you can only output Persian words in the text.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class, output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character.  Text: دلم اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر سرگرم خوندن شاهكاراي ادبي دنيا باشم ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم خوب شه و با اشتياق ب خوندن ادامه بدم <sup>a</sup>
assistant	1
user	Then, list the top K most influential words that contributed to this classification in CSV format (in a single line). Make sure to provide only K Persian words that exist in the original text and don't output any other token.  Text: دلم اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر سرگرم خوندن شاهكاراي ادبي دنيا باشم ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم خوب شه و با اشتياق ب خوندن ادامه بدم <sup>a</sup>
assistant	دلم, سرگرم, خوب, اشتياق, ادامه <sup>b</sup>
system	You are an emotion classifier. You are provided with some influential words that have been extracted from the text. You must classify the emotion based only on these words. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character. Text:  Text: دلم, سرگرم, خوب, اشتياق, ادامه <sup>b</sup>
assistant	1
system	You are an emotion classifier. In the text, some influential words have been replaced with the placeholder [حذف شده] <sup>c</sup> . You must classify the emotion based on the text, considering these [حذف شده] <sup>c</sup> words as part of the context. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character.  Text: [حذف شده] اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر [حذف شده] خوندن شاهكاراي ادبي دنيا باشم: ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم [حذف شده] شه و با [حذف شده] ب خوندن [حذف شده] بدم <sup>d</sup>
assistant	0

<sup>a</sup> **Translation:** "My heart wants that feeling where my mind is so free, I'm so engrossed in reading the literary masterpieces of the world, and suddenly at 3 o'clock I look at the window and see snow falling, and my mood gets many times better, and I continue reading with enthusiasm." **Transliteration:** delam un hâli-o mikhâd ke onghadr fekrâ rahâ bâshe, onghadr sargarm-e khundan-e shâhkârâ-ye adabi-ye donyâ bâsham, sâ'at-e 3 yeho be panjare negâh konam, bebinam barf miyâd, va dah-hâ barâbar hâlam khub she, va bâ eshtiâq be khundan edâme bedam.

<sup>b</sup> **Translation:** edâme 'continuation', eshtiâq 'enthusiasm', khub 'good', sargarm 'engaged', delam 'my heart'.

<sup>c</sup> **Translation:** "Removed" **Explanation:** hazf shode.

<sup>d</sup> **Translation:** "[removed] wants that feeling where my mind is so free, I'm so [removed] in reading the literary masterpieces of the world, and suddenly at 3 o'clock I look at the window and see snow falling, and my mood gets many times [removed], and I continue reading with [removed]." **Transliteration:** [hazf shode] un hâli-o mikhâd ke onghadr fekrâ rahâ bâshe, onghadr [hazf shode] khundan-e shâhkârâ-ye adabi-ye donyâ bâsham, sâ'at-e 3 yeho be panjare negâh konam, bebinam barf miyâd, va dah-hâ barâbar hâlam [hazf shode] she, va bâ [hazf shode] be khundan [hazf shode] bedam.

Table 11: This table illustrates the complete interaction protocol between system, user, and assistant roles for a single data sample during evaluation. It covers the Predict-then-Explain (P-E) paradigm across four stages: (1) standard classification, (2) influential word extraction, and (3) classification using only influential words (4) classification using modified input with influential words removed.

Role	Content
system	You are an emotion classifier. You must classify the emotion and output the top influential words in CSV format. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable. For top influential words, you can only output Persian words in the text.
user	List the top K most influential words that contributed to this classification in CSV format (in a single line). Make sure to provide only K Persian words that exist in the original text and don't output any other token.  Text: دل‌م اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر سرگرم خوندن شاهكاراي ادبي دنيا باشم ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم خوب شه و با اشتياق ب خوندن ادامه بدم <sup>a</sup>
assistant	دل‌م, سرگرم, خوب, اشتياق, ادامه <sup>b</sup>
user	Then, Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class, output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character.  Text: دل‌م اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر سرگرم خوندن شاهكاراي ادبي دنيا باشم ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم خوب شه و با اشتياق ب خوندن ادامه بدم <sup>a</sup>
assistant	1
system	You are an emotion classifier. You are provided with some influential words that have been extracted from the text. You must classify the emotion based only on these words. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character. Text:  Text: دل‌م, سرگرم, خوب, اشتياق, ادامه <sup>b</sup>
assistant	1
system	You are an emotion classifier. In the text, some influential words have been replaced with the placeholder [حذف شده] <sup>c</sup> . You must classify the emotion based on the text, considering these [حذف شده] <sup>c</sup> words as part of the context. For classifying, you are strictly required to output only one of the following English numbers: 0, 1, 2, 3, 4, or 5. No other output is acceptable.
user	Classify the following text into one of the categories: 'Sadness':0, 'Happiness':1, 'Anger':2, 'Surprise':3, 'Hatred':4, 'Fear':5. For each class output the mapped number. Only output an English number showing the class of the text. Make sure not to output any other character.  Text: [حذف شده] اون حالو ميخواد كه اونقدر فكرم رها باشه اونقدر [حذف شده] خوندن شاهكاراي ادبي دنيا باشم ساعت ۳ يهو ب پنجره نگاه كنم ببينم برف مياد و ده ها برابر حالم [حذف شده] شه و با [حذف شده] ب خوندن [حذف شده] ادامه بدم <sup>d</sup>
assistant	0

<sup>a</sup> **Translation:** "My heart wants that feeling where my mind is so free, I'm so engrossed in reading the literary masterpieces of the world, and suddenly at 3 o'clock I look at the window and see snow falling, and my mood gets many times better, and I continue reading with enthusiasm." **Transliteration:** delam un hâli-o mikhâd ke onghadr fekrâ rahâ bâshe, onghadr sargarm-e khundan-e shâhkârâ-ye adabi-ye donyâ bâsham, sâ'at-e 3 yeho be panjare negâh konam, bebinam barf miyâd, va dah-hâ barâbar hâlam khub she, va bâ eshtiâq be khundan edâme bedam.

<sup>b</sup> **Translation:** edâme 'continuation', eshtiâq 'enthusiasm', khub 'good', sargarm 'engaged', delam 'my heart'.

<sup>c</sup> **Translation:** "Removed" **Explanation:** hazf shode.

<sup>d</sup> **Translation:** "[removed] wants that feeling where my mind is so free, I'm so [removed] in reading the literary masterpieces of the world, and suddenly at 3 o'clock I look at the window and see snow falling, and my mood gets many times [removed], and I continue reading with [removed]." **Transliteration:** [hazf shode] un hâli-o mikhâd ke onghadr fekrâ rahâ bâshe, onghadr [hazf shode] khundan-e shâhkârâ-ye adabi-ye donyâ bâsham, sâ'at-e 3 yeho be panjare negâh konam, bebinam barf miyâd, va dah-hâ barâbar hâlam [hazf shode] she, va bâ [hazf shode] be khundan [hazf shode] bedam.

Table 12: This table illustrates the complete interaction protocol between system, user, and assistant roles for a single data sample during evaluation. It covers the Explain-then-Predict (E-P) paradigm across four stages: (1) influential word extraction, (2) standard classification, and (3) classification using only influential words (4) classification using modified input with influential words removed.

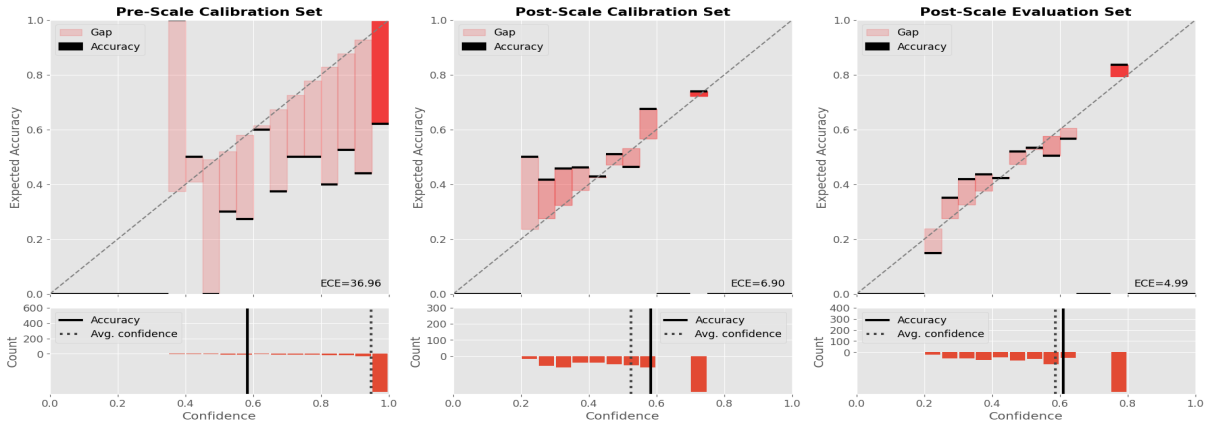


Figure 6: Reliability diagrams of the Chat GPT-4-turbo model under P-E paradigm.

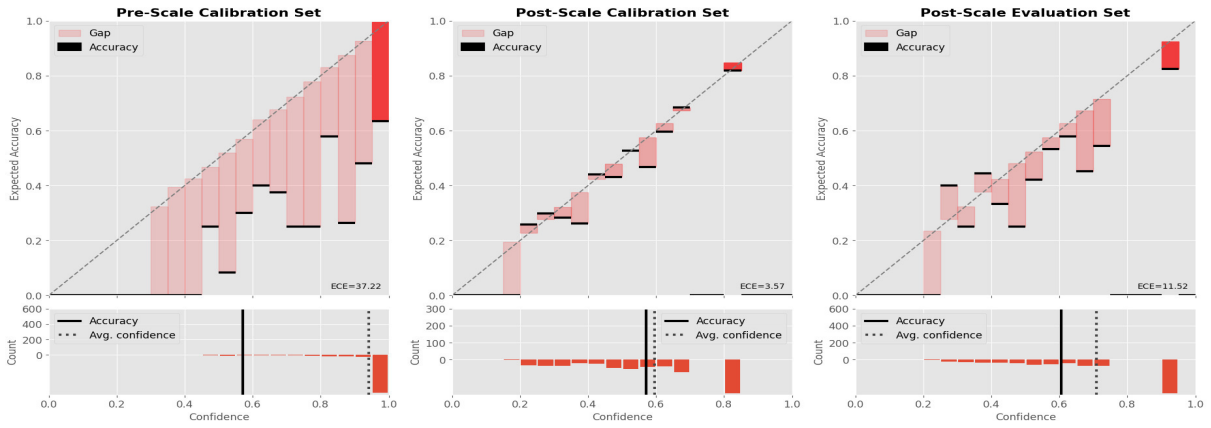


Figure 7: Reliability diagrams of the Chat GPT-4-turbo model under E-P paradigm.

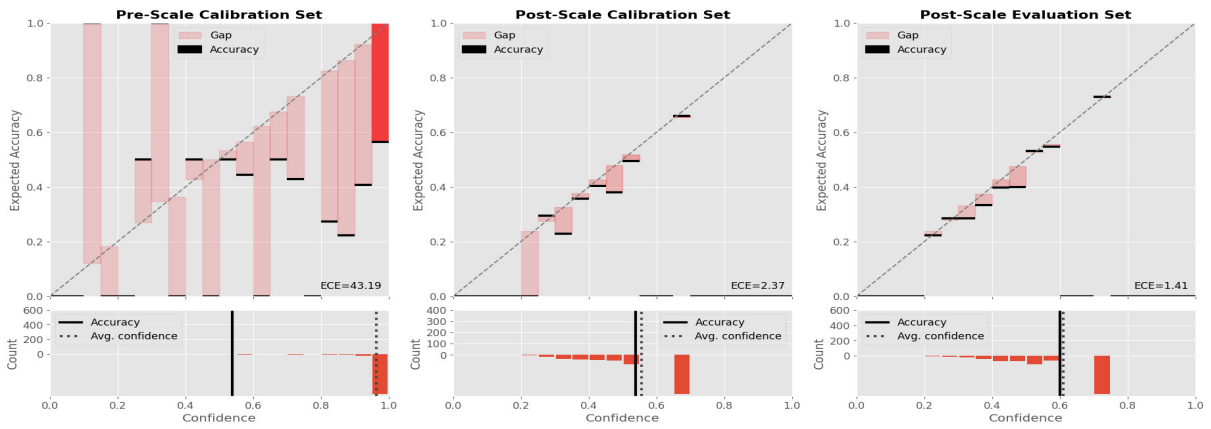


Figure 8: Reliability diagrams of the Chat Llama3.3-70B-instruct model under P-E paradigm.

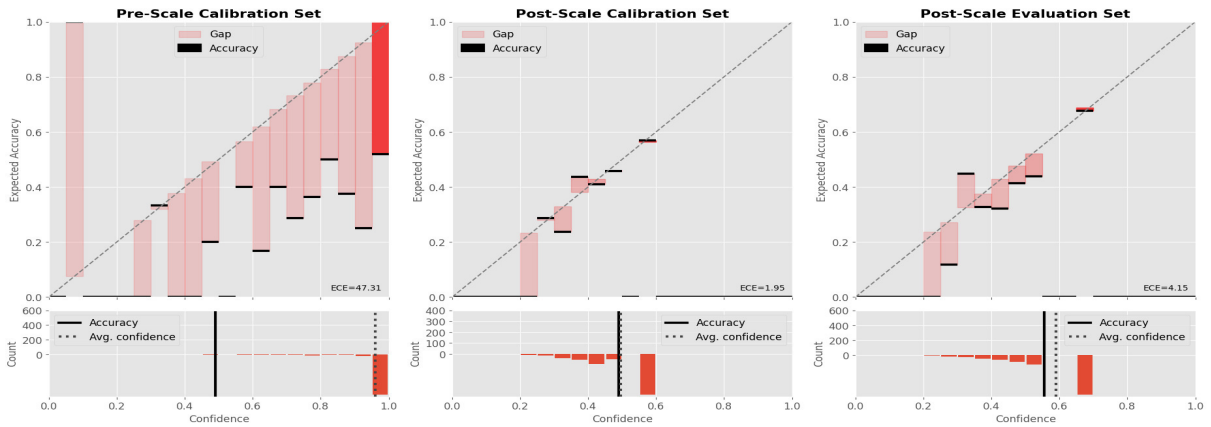


Figure 9: Reliability diagrams of the Chat Llama3.3-70B-instruct model under E-P paradigm.

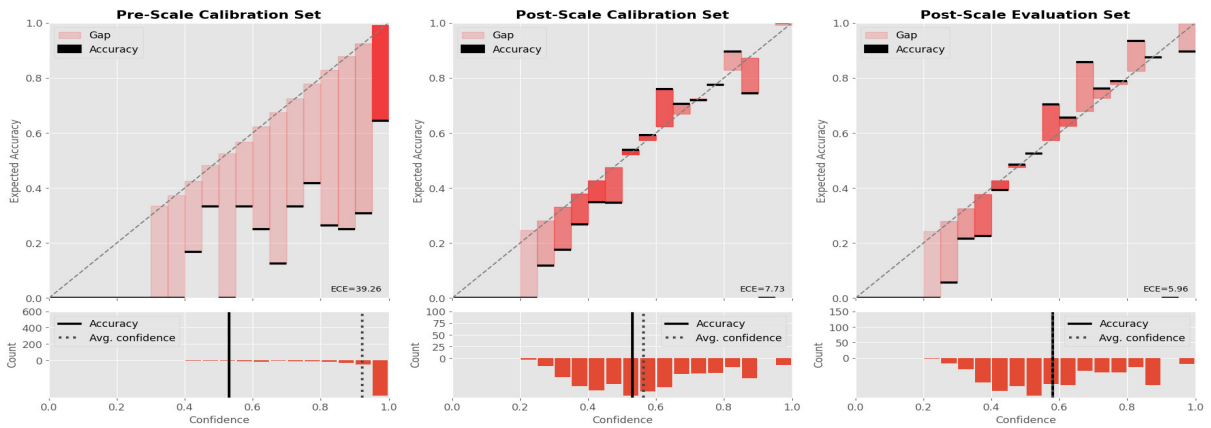


Figure 10: Reliability diagrams of the DeepSeek-V3 model under P-E paradigm.

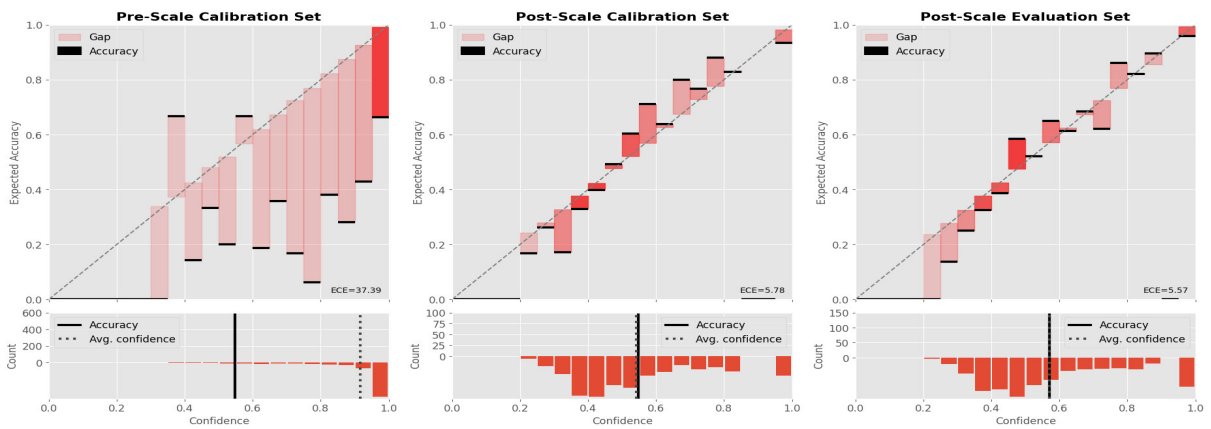


Figure 11: Reliability diagrams of the DeepSeek-V3 model under E-P paradigm.

Setting	Label	Model (Top-k Words)	Human (Top-k Words)
			من خر نیستم؛ دهن که باز می کنی می فهمم که آخرین کتابی که خوندی برمی گرده به چهار سالگیت، تازه اونم مامانت خونده برات نه خودت، پس برای من ادای روشن فکرای کتاب خون خفن رو در نیار که واقعاً چندشم میشه!!! نصف شبی گند زدی تو اعصابم!
PE	Anger	چندشم، گند، اعصابم، نیار، خر	خر، دهن، می فهمم، گند، اعصابم
EP	Anger	کتاب، مامانت، روشن، فکرای، کتابخون	خر، دهن، می فهمم، گند، اعصابم
			شهر خرم آبادم ، تو اتاق انتهایی خونه مادربزرگم در حالی که پنجره بازه ... صدای جیر جیرک و شاخ و برگ درختا میاد و هوای سرشار از اکسیژن اتاق رو پر کرده ، پاهام از زیر پتو بیرونه و با نسیم خنکی چند دقیقه یک بار ماساژ داده میشه
PE	Happiness	نسیم، خنکی، پنجره، بازه، اکسیژن	ماساژ، خنکی، سرشار، نسیم، اکسیژن
EP	Happiness	خرم، آبادم، پنجره، جیرک، نسیم	ماساژ، خنکی، سرشار، نسیم، اکسیژن
			#فرشته_حسینی ستاره این روزهای سینمای ایران، در #جشنواره_فیلم_فجر هم حضور دارد. #سیدرضامحمدی شاعر است و مقام نخست جشنواره #شعر فجر را به دست آورد. این دو هنرمند و شاعر اصالتاً اهل افغانستان هستند. موفقیت هر دو هموطن افغانستانی در فضای سینما و ادبیات ایران خوشایند است. بیش باد.
PE	Happiness	خوشایند، موفقیت، ستاره، شاعر، افغانستان	خوشایند، موفقیت، بیش باد، هموطن، ستاره
EP	Happiness	فرشته_حسینی، جشنواره_فیلم_فجر، سیدرضامحمدی، شعر فجر، افغانستان	خوشایند، موفقیت، بیش باد، هموطن، ستاره

Table 13: Representative Persian samples comparing GPT-4-turbo and human-identified influential words under Predict-then-Explain (P-E) and Explain-then-Predict (E-P) settings.

Setting	Label	Model (Top-k Words)	Human (Top-k Words)
			مشکل و بدبختی ایران همینجاست که تموم فسادا از زیر عمامه امثال تو میاد بیرون و برای فریب مردم گیر دادین به دو تار موی زنان... ریش شما ریشه مردم رو سوزونده... خدا خر رو شناخته که بهش شاخ نداده... این عوعو سگان شما هم نیز بگذرد... #انسانم_آرزوست
PE	ANGER	مشکل، بدبختی، فسادا، فریب، سوزونده	امثال، خر، سگ، فریب، بدبختی
EP	Hatred	مشکل، بدبختی، فسادا، عمامه، فریب	امثال، خر، سگ، فریب، بدبختی
			میتونم با نمودار و عدد و رقم بهتون نشون بدم اثر درمان دارویی روی روح و روان من باورنکردنیه. خودمم عمق داستان رو نمیدونستم تا وقتی که دو دسته دیتا از پارسال و امسال رو کنار هم گذاشتم و برگام ریخت. کاش زودتر به فکرش میفتم.
PE	Surprise	باورنکردنیه، داستان، دیتا، برگام، ریخت	باورنکردنی، برگام، نمیدونستم، عمق، اثر
EP	Surprise	نمودار، درمان، روح، روان، داستان	باورنکردنی، برگام، نمیدونستم، عمق، اثر
			#ما_نگین_کمانیها از ترس از دست دادن دوستان قدیمی و یه دنیا خاطره، از ترس انگشت نما شدن و از ترس جهل شما پارانویدها فقط بازیگران بهتری نسبت به سایرین هستیم. ما هیولا نیستیم. انسانیم و انتظاری جز انسانیت نداریم. #انسانم_آرزوست
PE	Fear	ترس، انگشت، جهل، هیولا	ترس، بازیگران، هیولا، پارانویدها، نگشت نما شدن
EP	Fear	ترس، دست، دنیا، جهل، انسانیت	ترس، بازیگران، هیولا، پارانویدها، نگشت نما شدن
			یادمه شبهای سرد زمستان غرب کشور از عراقیها نمیترسیدیم. از مجاهدین ضدخلق دائم در هراس بودیم چون فارسی حرف میزدن و به پشت سنگرها نفوذ میکردن و با سیم برنده و کلنگ سربازهای سر پست نگهبانی رو میکشتم. عملیات چلچراغ شاهد این ادعاست
PE	Fear	نمیترسیدیم، هراس، نفوذ، میکشتم، عملیات	هراس، میکشتم، مجاهدین، ضدخلق، سیم برنده
EP	Fear	شبهای، سرد، زمستان، هراس، عملیات	هراس، میکشتم، مجاهدین، ضدخلق، سیم برنده

Table 14: Representative Persian samples comparing GPT-4o and human-identified influential words under Predict-then-Explain (P-E) and Explain-then-Predict (E-P) settings.