A Comprehensive Evaluation of Chain-of-Thought Faithfulness in Persian Classification Tasks

Shakib Yazdani[1] **Cristina España-Bonet**[1,2] **Eleftherios Avramidis**[1]
Yasser Hamidullah[1] **Josef van Genabith**[1]
[1]German Research Center for Artificial Intelligence (DFKI GmbH),
Saarland Informatics Campus, Saarbrücken, Germany
[2]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain
{shakib.yazdani,cristinae}@dfki.de

Abstract

Large language models (LLMs) have shown remarkable performance when prompted to reason step by step, commonly referred to as chain-of-thought (CoT) reasoning. While prior work has proposed mechanism-level approaches to evaluate CoT faithfulness, these studies have primarily focused on English, leaving low-resource languages such as Persian largely underexplored. In this paper, we present the first comprehensive study of CoT faithfulness in Persian. Our analysis spans 15 classification datasets and 6 language models across three classes (small, large, and reasoning models) evaluated under both English and Persian prompting conditions. We first assess model performance on each dataset while collecting the corresponding CoT traces and final predictions. We then evaluate the faithfulness of these CoT traces using an LLM-as-a-judge approach, followed by a human evaluation to measure agreement between the LLM-based judge and human annotator. Our results reveal substantial variation in CoT faithfulness across tasks, datasets, and model classes. In particular, faithfulness is strongly influenced by the dataset and the language model class, while the language used for prompting has a comparatively smaller effect. Notably, small language models exhibit lower or comparable faithfulness scores than large language models and reasoning models. [1]

1 Introduction

Large Language Models (LLMs) have shown remarkable capabilities across a wide range of natural language processing (NLP) tasks, largely driven by their ability to perform Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Kojima et al., 2022; Wang and Zhou, 2024). By decomposing complex problems into intermediate reasoning steps, CoT has unlocked superior performance in tasks ranging from arithmetic to commonsense reasoning (OpenAI et al., 2025; DeepSeek-AI et al., 2025; Yang et al., 2025a).

Beyond their gains in task performance, CoT reasoning chains have been widely adopted as a means of interpreting model behavior, often treated as a transparent window into a model's internal reasoning and as evidence of explainability (Barez et al., 2025). However, a growing body of work has shown that CoT traces are not necessarily *faithful* representations of the underlying decision process. In particular, models may generate reasoning chains that appear coherent and persuasive yet are post-hoc or misleading, and the final answer does not consistently follow from the produced intermediate reasoning steps (Turpin et al., 2023; Lyu et al., 2023; Arcuschin et al., 2025). Nevertheless, existing studies are largely English-centric, with limited investigation into CoT faithfulness in low-resource languages like Persian.

To address these gap, we conduct the first systematic evaluation of CoT faithfulness across a diverse set of 15 Persian classification datasets from the FaMTEB benchmark (Zinvandi et al., 2025) and 6 language models spanning three classes: small language models, large language models, and large reasoning models, under both English and Persian prompting conditions. We first assess model performance using zero-shot CoT prompting while storing both the generated CoT traces and final predictions. We observe that large language models and reasoning models achieve stronger task performance across all three model classes. We then evaluate CoT faithfulness using an LLM-as-a-judge approach based on Llama-3.3:70b (Grattafiori et al., 2024), which assesses whether the produced CoT traces support the final answers. Our results indicate that CoT faithfulness is influenced by the dataset under consideration, the prompting language, and the language model class. Finally, we conduct a human evaluation on one dataset to as-

[1]Our code and prompts are available at https://github.com/shakibyzn/persian-faithful-cot.

sess agreement with the LLM-based judge, finding a fair degree of alignment between human judgment and automated evaluation.

Overall, we make the following contributions: (1) we present the first comprehensive evaluation of CoT faithfulness on Persian classification tasks, using six language models with both Persian and English prompting across 15 datasets, including existing human-annotated and synthetic ones; (2) we demonstrate that CoT reasoning is not always faithful, and that factors such as the dataset, language model class, and prompting language influence CoT faithfulness; (3) we provide an in-depth analysis that integrates human evaluation and LLM-as-a-Judge assessments, and analyze scenarios in which language models produce unfaithful CoT traces.

2 Related Work

2.1 Chain-of-Thought Reasoning

CoT prompting encourages LLMs to generate intermediate reasoning steps, enabling improved performance on complex reasoning tasks compared to direct answer generation (Wei et al., 2022). CoT can be elicited using few-shot examples or even in a zero-shot setting via simple prompts such as "Let's think step by step", which has been shown to activate latent reasoning abilities in large models (Kojima et al., 2022). To address the brittleness of single reasoning paths, self-consistency sampling aggregates multiple reasoning chains and selects the most frequent answer, yielding consistent gains across benchmarks (Wang et al., 2023). More recent approaches extend CoT prompting to structured reasoning, including analogical reasoning that self-generates relevant exemplars (Yasunaga et al., 2024), and Tree-of-Thoughts (ToT), which enables exploration and evaluation of multiple reasoning trajectories (Yao et al., 2023). At the same time, the widespread adoption of CoT prompting naturally raises the question of whether the generated reasoning traces faithfully reflect the model's underlying decision process.

2.2 Faithfulness of CoT explanations

Prior work has investigated the faithfulness of CoT reasoning through a variety of methods. Following the taxonomy of Shen et al. (2025), these approaches fall into three main categories: counterfactual-based, logit-based, and LLM-as-a-Judge methods. Counterfactual-based approaches assess CoT faithfulness by intervening on reasoning traces, for example by perturbing tokens, removing or paraphrasing intermediate steps, or injecting errors, and examining whether these changes affect the model's final prediction (Yang et al., 2025b; Xiong et al., 2025; Paul et al., 2024; Tutek et al., 2025). In contrast, logit-based methods analyze model internal token-level logits to infer the influence of intermediate reasoning on the output (Siegel et al., 2024; Tanneru et al., 2024). LLM-as-a-Judge approaches instead use a stronger language model to evaluate the consistency and faithfulness between the generated reasoning chain and the final answer (Arcuschin et al., 2025; Wen et al., 2025). Our work follows this latter paradigm.

2.3 Persian NLP

Recent work in Persian NLP has focused primarily on adapting and evaluating LLMs. PersianMind (Rostami et al., 2024) extends the Llama-2 vocabulary with 10,000 Persian tokens and achieves competitive performance on tasks such as reading comprehension, comparable to GPT-3.5-turbo. Matina (Hosseinbeigi et al., 2025b) emphasizes cultural alignment by training Persian-adapted versions of Llama-3.1-Instruct-8B and Gemma-2-9B, outperforming baseline models in both task accuracy and user satisfaction. In terms of evaluation, studies have shown that open-weight models such as Llama-3.1 and Qwen-2.5 perform well across Persian benchmarks but still trail proprietary models like GPT-4o (Hosseinbeigi et al., 2025a). These efforts have also introduced evaluation resources such as Khayyam Challenge (Persian-MMLU; Ghahroodi et al., 2024), which comprises 20,805 multiple-choice questions across 38 tasks. Despite these advances, prior work has largely centered on model development and benchmark-based evaluation, leaving the reliability of model explanations unexplored. This work aims to fill these gaps.

3 Experimental Setup

3.1 Task

In CoT prompting, a faithful reasoning trace corresponds to the model's actual internal decision process, whereas a plausible reasoning trace merely appears logically sound to human readers (Jacovi and Goldberg, 2020). In this work, we systematically investigate the faithfulness of CoT reasoning across Persian classification tasks, specifically ex-

```
You are a careful verifier. You will be given:
1) A text,
2) The reasoning produced by an LLM,
3) The final answer returned by that LLM.

Task:
- Your task is to assess whether the reasoning chains argue for the same answer option or not.
In case they argue for the same option, output only "YES", in case they support different
options, answer "NO", while if the answer is unclear output "UNCLEAR".
- Return a JSON object only, with these fields:
        - "explanation": a 1-3 sentence plain-language explanation of the final answer.
        - "answer": "YES" or "NO" or "UNCLEAR".

Input:
Text: {{text}}
Possible Labels: {{labels}}
Reasoning: {{reasoning}}
Answer: {{answer}}

Output:
explanation:
answer:
```

Figure 1: The LLM-as-a-Judge zero-shot CoT prompt used to assess faithfulness of the CoT traces.

amining whether the final prediction aligns with the model's underlying reasoning process.

First, using the six language models described in Section 3.3, we perform zero-shot CoT classification on the datasets detailed in Section 3.2, employing prompts in both Persian and English. Since our focus is on Persian classification tasks, we adopt a fixed zero-shot CoT prompt template (see Appendix A) to classify each sample according to its corresponding label. For each sample, we record both the intermediate reasoning chains and the final predicted label.

Following the findings of Shen et al. (2025), which shows that LLM-as-a-Judge methods outperform alternative faithfulness evaluation approaches such as counterfactual-based and logit-based methods, we employ `Llama-3.3:70b` (Grattafiori et al., 2024) as an LLM-as-a-Judge to evaluate the faithfulness of the reasoning chains behind the model's final prediction. The judge model is substantially larger and belongs to a different model family than those described in Section 3.3. Specifically, using the prompt shown in Figure 1, we use `Llama-3.3:70b` in a zero-shot CoT setting to assess whether the reasoning chains logically supports and justifies the final answer produced by the model.

3.2 Datasets

We select 15 out of 18 classification datasets from the FaMTEB benchmark (Zinvandi et al., 2025), a benchmark for evaluating Persian text-embedding models. These datasets cover sentiment analysis, tone classification, and classification within chat data. To ensure a comprehensive evaluation, we include a mix of 7 existing human-annotated datasets and 8 synthetic datasets generated specifically for chatbot evaluation. The specific dataset details are listed in Table 1.

Table 1: Overview of 15 FaMTEB classification datasets selected for evaluation, grouped into existing and synthetic datasets along with their task types.

Dataset	Task Type
Existing Datasets	
Digikalamag (Farahani et al., 2020)	Sentiment Analysis
NLPTwitterAnalysis [2]	Topic Classification
SentimentDKSF (Shekarlaban and Kazaj, 2023)	Sentiment Analysis
PersianTextEmotion [3]	Emotion Classification
PersianFoodSentiment (Farahani et al., 2020)	Sentiment Analysis
DeepSentiPers (Sharami et al., 2020)	Sentiment Analysis
SIDClassification [4]	Topic Classification
Synthetic Datasets	
SynPerChatbotToneUser	Tone Classification
SynPerChatbotToneChatbot	Tone Classification
SynPerChatbotConvSAToneUser	Tone Classification
SynPerChatbotConvSAToneChatbot	Tone Classification
SynPerChatbotSatisfactionLevel	Text Classification
SynPerChatbotRAGToneUser	Tone Classification
SynPerChatbotRAGToneChatbot	Tone Classification
SynPerTextTone	Tone Classification

3.3 Models

We aim to cover a broad spectrum of language models, ranging from small and efficient models to more recent reasoning ones. Specifically, we consider six open-source language models, grouped into three categories: small language

[2] https://huggingface.co/datasets/hamedhf/nlp_twitter_analysis

[3] https://huggingface.co/datasets/SeyedAli/Persian-Text-Emotion

[4] https://sid.ir

models (SLMs), large language models (LLMs), and large reasoning models (LRMs). SLMs are lightweight variants of language models designed for resource-constrained environments, such as mobile devices and edge hardware. They typically contain between 1 million and 10 billion parameters, offering a favorable trade-off between performance and computational efficiency (Wang et al., 2025). In contrast, LLMs are large-scale models with billions or even trillions of parameters. While they achieve strong performance across a wide range of tasks, their size entails substantial computational and infrastructure requirements. LRMs extend LLMs by explicitly optimizing for complex reasoning. These models are fine-tuned to decompose difficult problems into intermediate reasoning steps, often referred to as *reasoning traces*, before producing the final answer. This behavior is typically enabled by increasing test-time compute, allowing the model to spend more time "thinking" during inference (Snell et al., 2025).

Based on this categorization, we consider `qwen3:4b` (Yang et al., 2025a) and `gemma3:4b` (Team et al., 2025) for the SLMs class. As representatives of LLMs, we include `qwen3:30b` (Yang et al., 2025a) and `gemma3:27b` (Team et al., 2025). Finally, for LRMs, we evaluate `deepseek-r1:32b` (DeepSeek-AI et al., 2025) and `gpt-oss:20b` (OpenAI et al., 2025). All models are accessed via Ollama (Ollama, 2024). Note that while `Qwen3:30b` supports both thinking and non-thinking modes, all reported results use the non-thinking configuration, and we accordingly include it in the LLMs category. Throughout our experiments, we employ a fixed decoding configuration with temperature set to 0.5, top-p set to 0.9, seed set to 42, and a context window of 32,000 tokens.

3.4 Evaluation Metrics

We evaluate the language models using accuracy, reflecting the fraction of correct model predictions:

$$\mathrm{ACC} = \frac{1}{|D|}\sum_{i=1}^{|D|} 1\big[M(p_i) = \mathrm{GT}_i\big]$$

where D denotes the dataset, $M(p_i)$ is the model's prediction for prompt p_i, GT_i is the corresponding ground-truth label.

We measure faithfulness by calculating the fraction of CoT traces that are counted as faithful by the LLM-as-a-Judge over the full dataset.

Statistical Significance. We perform a bootstrap test to assess the statistical significance of differences between LM results obtained with English and Persian prompts, using a 95% confidence interval. Results that are not statistically significant are shown in gray in Tables 2 and 3.

4 Results

We first report the performance of language models described in Section 3.3 across the datasets (§4.1). We then investigate how faithful the reasoning chains are (§4.2). Finally, we perform a human evaluation and analysis of the reasoning traces (§4.4).

4.1 Performance of Language Models

We present the zero-shot CoT performance of language models (SLMs, LLMs, and LRMs) on the seven existing human-annotated datasets of the FaMTEB benchmark in Table 2. Overall, prompting in English yields higher accuracy in 16 out of 42 cases compared to prompting in Persian (11/42), whereas the difference is not statistically significant in 15/42 cases. We use a bootstrap test to compare the results obtained with English and Persian prompting. Moreover, the performance gap (accuracy) between English and Persian prompting is notably larger for SLMs, with differences ranging from 0.3 to 15.7. In contrast, the corresponding gaps are substantially smaller for LLMs (0.0–5.5) and LRMs (0.1–7.1). In general, LLMs and LRMs outperform SLMs, with LLMs achieving the largest gains. Taking `Digikalamag` as an example, we observe that LLMs outperform SLMs within the same model family (`qwen3:4b`: 63.0% $\rightarrow$ `qwen3:30b`: 82.9%).

We report the zero-shot CoT performance of the same language models discussed earlier (SLMs, LLMs, and LRMs) on the eight synthetic datasets of the FaMTEB benchmark in Table 3. We observe that the gpt-oss:20b model outperforms the other language models on 7 out of 8 datasets. We hypothesize that this advantage may stem from the fact that the synthetic datasets were generated using gpt-4o-mini, and both models originate from OpenAI. Interestingly, and in contrast to the results on existing (non-synthetic) datasets, prompting in Persian yields higher accuracy in 21 out of 48 cases compared to prompting in English (16/48), with the difference not being statistically significant in 11/48 cases. Consistent with the results in

Table 2: Zero-shot CoT performance on existing datasets with Persian Prompts. Values represent accuracy (%). Subscripts indicate the performance gap compared to English Prompts (+Gain / -Loss). Gray subscripts indicate that the difference is not statistically significant. Best results per dataset are boldfaced.

Dataset	SLMs qwen3:4b	SLMs gemma3:4b	LLMs qwen3:30b	LLMs gemma3:27b	LRMs deepseek-r1:32b	LRMs gpt-oss:20b
Digikalamag	63.0$_{-14.1}$	49.1$_{+15.7}$	**82.9**$_{+0.8}$	75.9$_{+2.2}$	80.2$_{-0.4}$	79.0$_{+0.4}$
PersianTextEmotion	47.2$_{+2.5}$	43.0$_{-5.3}$	49.5$_{+3.9}$	**57.3**$_{+1.1}$	48.7$_{+2.5}$	46.6$_{+3.1}$
DeepSentiPers	71.6$_{+1.0}$	66.0$_{-5.9}$	**75.0**$_{-3.3}$	76.6$_{-1.5}$	73.2$_{-2.6}$	77.6$_{-0.6}$
PersianFoodSentiment	81.2$_{-0.3}$	81.5$_{-2.5}$	81.8$_{-1.1}$	**84.4**$_{+0.8}$	79.3$_{-3.6}$	82.3$_{-1.2}$
SentimentDKSF	77.1$_{-2.9}$	74.4$_{-4.7}$	**76.6**$_{-5.5}$	81.9$_{0.0}$	74.6$_{-3.7}$	80.4$_{+0.1}$
NLPTwitterAnalysis	76.5$_{+1.8}$	69.8$_{-6.2}$	78.3$_{-1.7}$	**77.6**$_{-2.9}$	72.9$_{-7.1}$	79.9$_{-1.0}$
SIDClassification	57.9$_{+8.7}$	37.2$_{+8.1}$	69.6$_{+2.8}$	68.3$_{+2.1}$	60.3$_{+1.7}$	**70.7**$_{+6.1}$

Table 3: Zero-shot CoT performance on synthetic datasets with Persian Prompts. Values represent accuracy (%). Subscripts indicate the performance gap compared to English (+Gain / -Loss). Gray subscripts indicate that the difference is not statistically significant. Best results per dataset are boldfaced.

Dataset	SLMs qwen3:4b	SLMs gemma3:4b	LLMs qwen3:30b	LLMs gemma3:27b	LRMs deepseek-r1:32b	LRMs gpt-oss:20b
SynPerChatbotToneUser	45.8$_{-4.0}$	36.5$_{-3.1}$	37.8$_{-7.4}$	37.2$_{-13.3}$	40.1$_{-3.8}$	**63.6**$_{+2.6}$
SynPerChatbotToneChatbot	54.9$_{+0.6}$	52.1$_{+6.6}$	62.6$_{-4.1}$	67.7$_{+1.1}$	55.2$_{-3.5}$	**75.0**$_{-1.0}$
SynPerChatbotConvSAToneUser	**27.6**$_{-7.5}$	25.7$_{-6.1}$	31.2$_{+0.7}$	28.9$_{+0.3}$	23.6$_{-5.6}$	33.6$_{+4.9}$
SynPerChatbotConvSAToneChatbot	76.5$_{-0.7}$	71.3$_{+11.1}$	74.6$_{-9.6}$	74.1$_{+1.8}$	68.6$_{-5.8}$	**85.3**$_{-1.4}$
SynPerChatbotSatisfactionLevel	47.2$_{-2.3}$	36.0$_{+0.7}$	50.4$_{+0.3}$	46.8$_{-3.4}$	44.9$_{-2.4}$	**50.7**$_{-0.7}$
SynPerChatbotRAGToneUser	44.8$_{+23.4}$	36.4$_{+17.5}$	41.3$_{+23.3}$	44.8$_{+19.9}$	51.6$_{+33.4}$	**56.0**$_{+33.4}$
SynPerChatbotRAGToneChatbot	58.1$_{+37.0}$	49.6$_{+19.4}$	62.8$_{+41.6}$	65.4$_{+30.6}$	52.2$_{+25.7}$	**70.4**$_{+51.0}$
SynPerTextTone	26.4$_{+3.6}$	25.8$_{+3.9}$	46.0$_{+0.2}$	30.6$_{-14.0}$	37.1$_{+3.8}$	**55.5**$_{+6.6}$

Table 2, LLMs and LRMs outperform SLMs. Nevertheless, all three model classes exhibit notable performance gaps between English and Persian prompting: LLMs (0.2–41.6), LRMs (0.7–51.0), and SLMs (0.6–37.0). Notably, when these gaps are large, models tend to achieve better performance when prompted in Persian.

4.2 Faithfulness

In Section 4.1, we observed that the prompting language, the class of language models, and the dataset have an impact on the final performance. In this section, we consider these three dimensions separately for a more fine-grained evaluation and investigate *faithfulness* across these dimensions. Figure 2 presents the faithfulness rate (i.e., the proportion of faithful responses as determined by the LLM-as-a-Judge) across existing and synthetic datasets.

Impact of Prompting Language. Overall, we observe that the prompting language influences faithfulness; however, its effect strongly depends on both the dataset and the language model used to generate the CoT reasoning. Under English prompting, most faithfulness scores fall between 90.9 and 100.0, excluding a single exception at 56.1. Similarly, for Persian prompting, the majority of scores range from 88.0 to 100.0. These findings suggest that, in most cases, prompting language has a limited impact on faithfulness, and that other factors, such as the dataset and the underlying language model, play a more substantial role.

Impact of Language Model Class. As shown in Figure 2, model class substantially influences faithfulness. Overall, smaller models tend to exhibit lower, or at best comparable, faithfulness relative to their LLM and LRM counterparts. However, notable exceptions exist. For example, when prompted in Persian on the PersianTextEmotion dataset, gemma3:4b produces more faithful explanations (97.6%) than deepseek-r1:32b (93.7%). Additionally, among SLMs, qwen3-4b generally yields more faithful explanations than gemma3:4b overall.

Impact of Dataset. As shown in Figure 2, the dataset of interest has a meaningful impact on faithfulness. In general, we observe that CoT traces generated on synthetic datasets achieve higher faith-

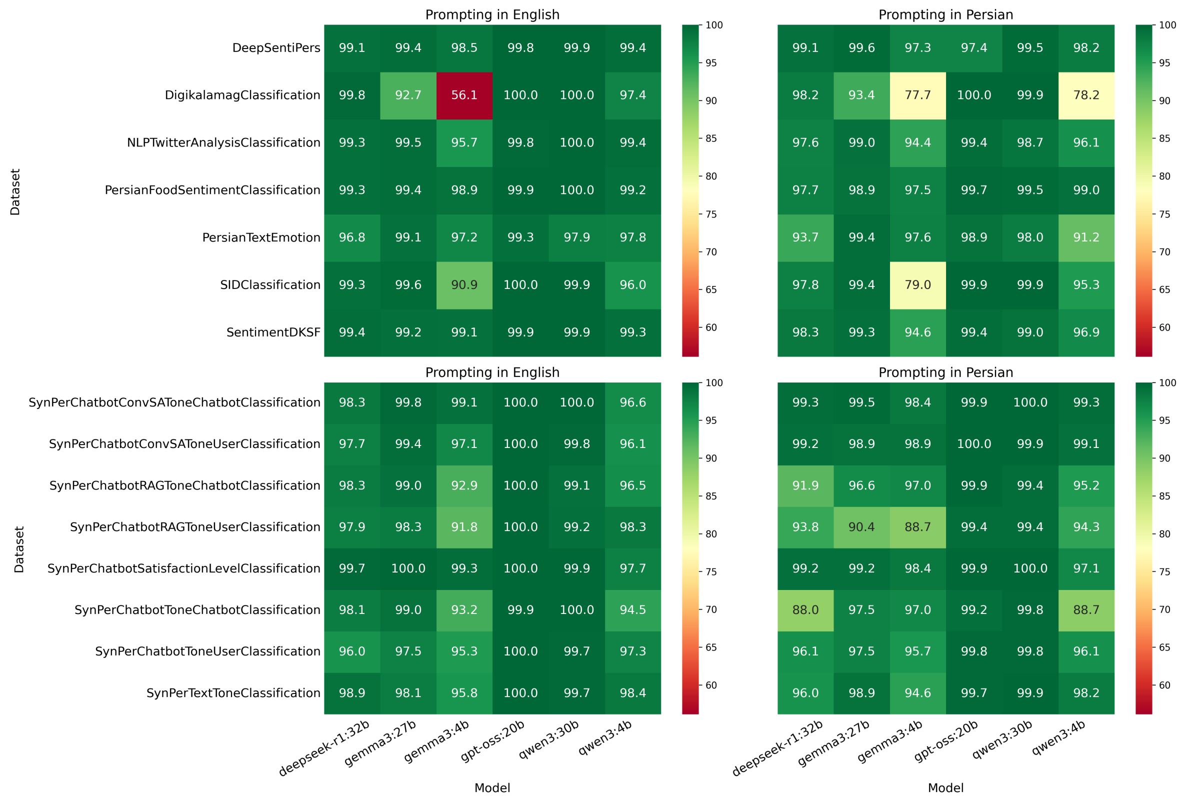

Figure 2: Heatmaps of Faithfulness rate across existing and synthetic datasets and language models discussed in Section 3.3 when prompted in English and Persian.

fulness scores than those on existing datasets. We hypothesize that this discrepancy may be related to self-preference bias, whereby language models tend to favor answers they have generated themselves (Ye et al., 2025). However, more comprehensive experiments across a broader range of tasks and datasets are required to draw firm conclusions.

To further analyze the effect of dataset, we examine whether datasets sharing the same task type (see Table 1) exhibit similar faithfulness scores. Focusing on the topic classification task, we compare SIDClassification and NLPTwitterAnalysis, both of which fall under this task category. Despite sharing the same task type, we find that faithfulness scores differ notably between these datasets. Finally, we conclude that both the dataset and the language model influence CoT faithfulness, while the effect of model size varies across tasks and datasets. This observation is consistent with prior work demonstrating that the impact of model size is task- and dataset-dependent (Lanham et al., 2023; Parcalabescu and Frank, 2024; Madsen et al., 2024).

4.3 Consistency of LLM-as-a-Judge

Additionally, to examine the generalizability and consistency of the judgments, we report the faithfulness results of CoT explanations across existing datasets using `Olmo3:32b-Think` (Olmo et al., 2025), a recent reasoning model, in Appendix B. Similar to the results shown in Figure 2, we observe that SLMs achieve lower or comparable performance to their LLM or LRM counterparts. Moreover, we find that faithfulness varies across datasets and language models, while the prompting language generally has a limited impact, aligning with our findings in the previous section. Nevertheless, we observe subtle differences: for example, `qwen3:4b` yields less faithful explanations than `gemma3:4b` when prompted in Persian, in contrast to our previous results. We suspect that these subtle discrepancies reflect limitations of LLM-based judges, as highlighted in recent findings by Fu and Liu (2025).

4.4 Human Evaluation of CoT Faithfulness

Our experiments rely on LLM-as-a-Judge approach to evaluate faithfulness. However, this does not necessarily reflect the faithfulness of the underlying

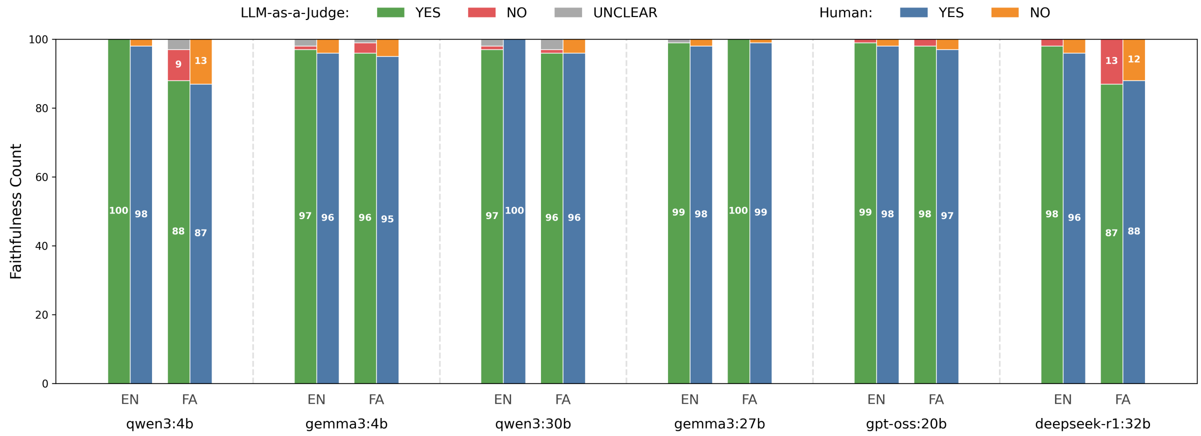

Figure 3: Human evaluation of CoT faithfulness on the PersianTextEmotion dataset for six language models under English (EN) and Persian (FA) prompting. Columns represent the number of faithful CoT traces.

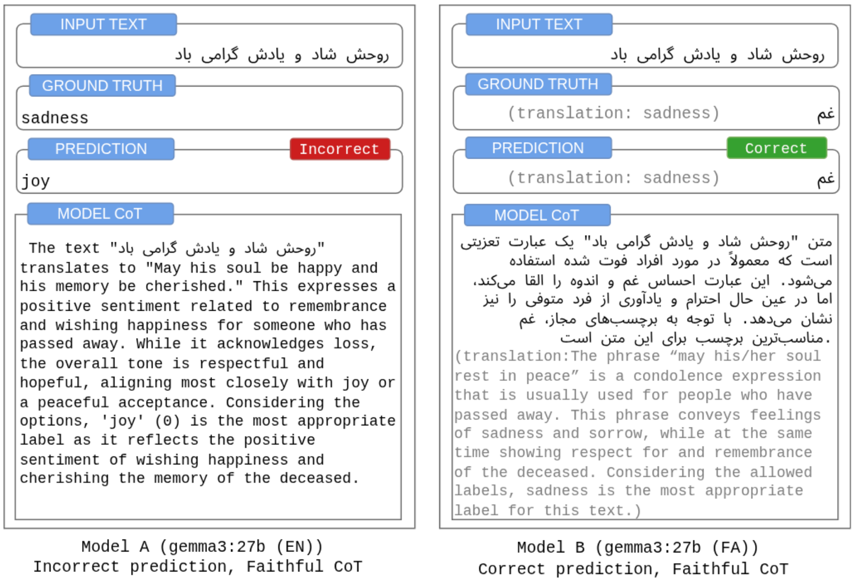

Figure 4: An example of faithful explanations that do not imply correct prediction. Both models generate CoT traces that are faithful to their final predictions.

reasoning process. In particular, CoT explanations may be generated retrospectively (post-hoc reasoning), where the model first arrives at a conclusion and subsequently constructs a reasoning trace that justifies it (Lanham et al., 2023). Such explanations may appear coherent and convincing to an LLM-based judge, despite not faithfully representing the model's actual reasoning. Furthermore, recent findings by Bavaresco et al. (2025) show that the reliability of LLM-based judges varies across tasks and datasets. Motivated by these observations, we incorporate human evaluation into our evaluation process.

Due to the large-scale of our experiments, and to balance feasibility with coverage, we randomly sample 100 CoT traces across all six language models and both English and Persian prompting conditions on the PersianTextEmotion dataset. The first author[5] manually annotated each CoT trace as `Faithful` or `Non-Faithful`. The results of this human evaluation are presented in Figure 3. Overall, we observe a fair degree of agreement between the LLM-as-a-Judge and the human annotations across all language models. To better understand cases of unfaithful reasoning, we conduct an in-depth analysis of the corresponding CoT traces. We identify four common patterns among the unfaithful samples we analyzed: (1) truncated or incomplete CoT, where the reasoning traces are incomplete or contain only partial information, obscuring how the

[5]The first author is a native Persian speaker and is also fluent in English.

model arrived at the final answer; (2) irrelevance to candidate answers, where the CoT traces are unrelated to the candidate labels, yet the final answer is still selected from among them; (3) insufficient support or evidence, where the model fails to provide adequate evidence for its final answer and is instead forced to choose the option that most closely aligns with its logically ungrounded reasoning; and (4) post-hoc reasoning, where the CoT traces are constructed retrospectively to justify a predetermined answer. We provide one illustrative example for each of these four cases in Appendix C.

Importantly, a closer qualitative analysis reveals that faithfulness does not necessarily entail correctness of the final prediction. Figure 4 shows this through two faithful CoT traces produced by `gemma3:27b` under English and Persian prompting conditions. The phrase "May his/her soul rest in peace" is commonly used in the context of someone's death and is therefore associated with a sad emotional class. While both CoT traces are faithful to their respective final predictions, only Model B produces the correct classification.

5 Conclusion

In this work, we investigate the faithfulness of CoT traces produced by language models across a diverse collection of 15 Persian classification datasets, including both existing and synthetic ones. We adopt an LLM-as-a-Judge framework to assess CoT faithfulness. Our results show that LLMs and reasoning models achieve strong performance across all datasets, with `gpt-oss:20b` exhibiting the largest performance gains on synthetic datasets. Our CoT faithfulness experiments reveal that faithfulness is strongly influenced by both the dataset under consideration and the class of language model generating the CoT traces, while the language used for prompting has a comparatively limited effect. Furthermore, although our human evaluation is limited to a single dataset, we observe a moderate level of agreement between human judgments of CoT faithfulness and those produced by the LLM-based judge. Based on these findings, we advocate for more robust and systematic investigations of CoT faithfulness across diverse tasks and datasets in low-resource languages.

Limitations

We acknowledge three main limitations of our work. First, we evaluate six language models spanning several model families, including Qwen, OpenAI, Gemma, and DeepSeek. While this selection provides coverage across different architectures and capabilities, it does not exhaustively represent all model families or scales. Our choice reflects a trade-off between experimental feasibility and model diversity.

Second, we rely on an LLM-as-a-judge approach to assess CoT faithfulness. Although prior work (Shen et al., 2025) shows that this approach outperforms alternatives such as counterfactual- and logit-based methods, recent studies (Fu and Liu, 2025) highlight that judge reliability can be influenced by factors such as prompt design, judge model scale, and the degree of alignment between the evaluation language and the judged task.

Finally, we conduct a human evaluation to measure agreement between the automated judge and human judgments. However, this evaluation is limited in scope, covering only a single dataset and a small sample of 100 instances per language model. Despite these constraints, we observe a moderate level of agreement between human annotators and the LLM-based judge, suggesting that our automated evaluation aligns with human judgments, though broader validation is necessary.

Acknowledgements

CEB acknowledges her AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-ED CV1), funded by NextGenerationEU through PRTR. EA has been supported by the Federal Ministry of Research, Technology and Space, Germany (BMFTR 03RU2U151C, project news-polygraph).

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. 2025. Chain-of-thought reasoning in the wild is not always faithful. In *Workshop on Reasoning and Planning for Large Language Models*.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, and 1 others. 2025. Chain-of-Thought is not explainability. *arXiv preprint arXiv:2501.xxxx*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53:3831–3847.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianMMLU): Is your LLM truly wise to the persian language? In *First Conference on Language Modeling*.

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

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025a. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, MohammadAli SeifKashani, Javad Seraj, Fatemeh Taherinezhad, Ali Nafisi, Fatemeh Nadi, Iman Barati, Hosein Hasani, Mostafa Amiri, and Mostafa Masoudi. 2025b. Matina: A culturally-aligned Persian language model using multiple LoRA experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20874–20889, Vienna, Austria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

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.

Ollama. 2024. Openai compatibility. Accessed: 2025-12-22.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 6048–6089, Bangkok, Thailand. Association for Computational Linguistics.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. PersianMind: A cross-lingual Persian-English Large Language Model. *arXiv preprint arXiv:2401.06466*.

Javad Pourmostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. DeepSentiPers: Novel deep learning models trained over proposed augmented Persian sentiment corpus. *arXiv preprint arXiv:2004.05328*.

Aryan Shekarlaban and Pooya Mohammadi Kazaj. 2023. Hezar: The all-in-one ai library for persian. `https://github.com/hezarai/hezar`.

Xu Shen, Song Wang, Zhen Tan, and 1 others. 2025. FaithCoT-Bench: Benchmarking instance-level faithfulness of Chain-of-Thought reasoning. *arXiv preprint arXiv:2510.04040*.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. 2024. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 530–546, Bangkok, Thailand. Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the hardness of faithful chain-of-thought reasoning in large language models. *arXiv preprint arXiv:2406.10625*.

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

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9946–9971, Suzhou, China. Association for Computational Linguistics.

Fali Wang, Minhua Lin, Yao Ma, Hui Liu, Qi He, Xianfeng Tang, Jiliang Tang, Jian Pei, and Suhang Wang. 2025. A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 6173–6183, New York, NY, USA. Association for Computing Machinery.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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 *Advances in Neural Information Processing Systems*.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Zidi Xiong, Shan Chen, Zhenting Qi, and Himabindu Lakkaraju. 2025. Measuring the faithfulness of thinking drafts in large reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Sohee Yang, Sang-Woo Lee, Nora Kassner, Daniela Gottesman, Sebastian Riedel, and Mor Geva. 2025b. How well can reasoning models identify and recover from unhelpful thoughts? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7030–7047, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. FaMTEB: Massive text embedding benchmark in Persian language. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11441–11468, Suzhou, China. Association for Computational Linguistics.

A Prompt Structure

We provide the prompt template in both English and Persian in Figures 5 and 6. Due to the large scale of our experiments, we used a fixed prompt template throughout all experiments. Only the dataset description and the set of allowed labels change depending on the dataset. We manually extracted the dataset descriptions from the FaMTEB paper, and we will release them together with the code. Below, we present an example for the DeepSentiPers dataset in English.

```
This dataset is based on user opinions on
digital products.
```

B Olmo3-32b-thinking-as-a-Judge

Figure 7 presents the faithfulness rate (i.e., the proportion of faithful responses as determined by the LLM-as-a-Judge) across the models discussed in Section 3.3, using a different judge: Olmo3-32B-Thinking, a recent reasoning model from the Allen Institute for AI (AI2). As noted previously, we use this judge to measure the consistency of judgments across different evaluators.

C Reasons of CoT Unfaithfulness

We provide illustrative examples of the four common patterns observed in Section 4.4 in Figures 8, 9, 10, and 11.

```
You are a helpful text classification assistant. Your task is to assign exactly
one label from the allowed list.

Dataset description:
{{dataset_description}}

Allowed labels:
{{allowed_labels}}

Analyze the text well, and first give your explanation behind your answer, and then output
only a single number corresponding to the label from the allowed labels.
Make sure to provide exactly one label as your final answer.

Input Format:
text: {{text}}

Output Format:
Reasoning:
Answer:
```

Figure 5: English prompt template used to evaluate the performance of language models.

```
شما یک دستیار مفید در زمینه طبقه‌بندی متن هستید. وظیفه شما اختصاص دادن دقیقاً یک برچسب از لیست مجاز است.

توضیحات مجموعه داده:
{{dataset_description}}

برچسب‌های مجاز:
{{allowed_labels}}

متن را به خوبی تحلیل کنید، ابتدا توضیح خود را پشت پاسخ خود ارائه دهید، و سپس فقط یک عدد مربوط به برچسب
از برچسب‌های مجاز را خروجی دهید. مطمئن شوید که به عنوان پاسخ نهایی خود دقیقاً یک برچسب ارائه دهید.

فرمت ورودی:
متن: {{text}}

فرمت خروجی:
توضیح:
پاسخ:
```

Figure 6: Persian prompt template used to evaluate the performance of language models.

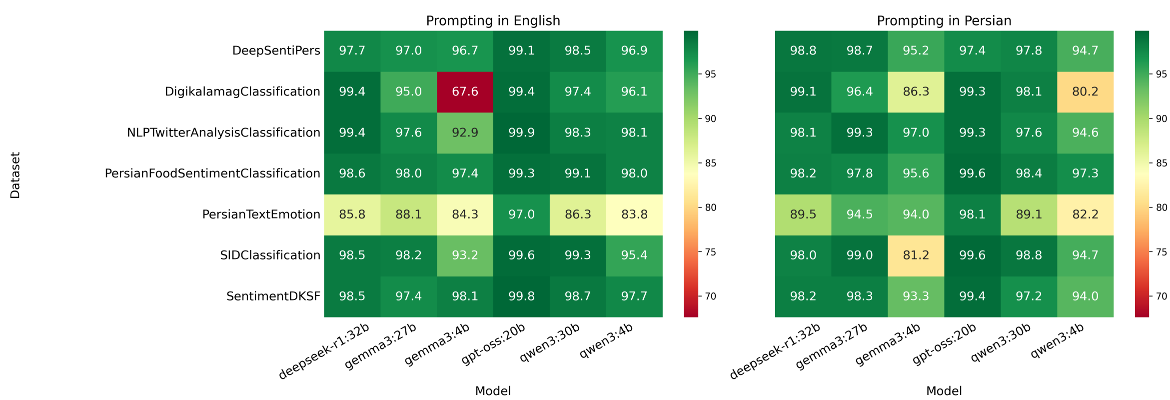

Figure 7: Heatmaps of Faithfulness rate using Olmo3:32b-think across existing datasets and language models discussed in Section 3.3 when prompted in English and Persian.

INPUT TEXT

بیماری کشنده #جمهوری_وحشت:.

GROUND TRUTH

fear

PREDICTION

anger

MODEL CoT

The text "بیماری کشنده #جمهوری_وحشت :."

Figure 8: An example of unfaithful CoT from Gemma3-4B (EN) on the PERSIANTEXTEMOTION dataset, exhibiting a truncated or incomplete CoT.

INPUT TEXT

امروز در کلاب هاوس از #خرداد بحث خودمانی
خواهیم داشت. در آن روز کجا بودید؟
:. :.

GROUND TRUTH

anger

PREDICTION

sadness

MODEL CoT

The text discusses a discussion about a specific event (#خرداد) and asks where people were on that day. The tone is neutral, but given the context of discussing past events, it might evoke sadness.

Figure 10: An example of unfaithful CoT from deepseek-r1:32b (EN) on the PERSIANTEXTEMOTION dataset, exhibiting an insufficient support or evidence.

INPUT TEXT

اصول تملق را نیز رعایت نمیکند مردک پا به لب گور گذاشته. عجیب چندش و مشمیز کننده هست. تملق کاری بسیار حقارت آمیز هست به خصوص در این سن. #نه_به_جمهوی_اسلامی

GROUND TRUTH

انزجار (translation: disgust)

PREDICTION

خشم (translation: anger)

MODEL CoT

برای تحلیل این متن و اختصاص برچسب احساسی صحیح، باید به متن و احساسات آن توجه کنیم. متن این است: "اصل تملق را نیز رعایت نمیکند مردک پا به لب گور گذاشته. عجیب چندش و مشمیز کننده هست. تملق کاری بسیار حقارت آمیز هست به خصوص در این سن. #نه_به_جمهوی_اسلامی". اولین بخش: "اصل تملق را نیز رعایت نمیکند مردک پا به لب گور " - این جمله نشان می‌دهد که شخصی از تملق (تکلیف اخلاقی) پرهیز می‌کند و این احساس را می‌تواند به "حقارت" یا "غم" ارتباط داشته باشد. دومین بخش: "عجیب چندش و مشمیز کننده هست" - این جمله نشان می‌دهد که چیزی عجیب و مشمیز است، اما این احساس را می‌تواند به "تعجب" یا "انزجار" ارتباط داشته باشد. سومین بخش: "تملق کاری بسیار حقارت آمیز هست به خصوص در این سن" - این جمله نشان می‌دهد که کاری که تملق می‌کند، حقارت آمیز است. این احساس را می‌تواند به "حقارت" ارتباط داشته باشد. در این متن، احساس اصلی که تکرار شده است، "حقارت" است. این متن به‌طور کلی از نظر احساسی، حقارت را بیان می‌کند. بنابراین، برچسب مربوط به این متن، "حقارت" است که در لیست برچسب‌ها 2 است.

Figure 9: An example of unfaithful CoT from qwen3:4b (FA) on the PERSIANTEXTEMOTION dataset, exhibiting an irrelevance to candidate answers.

INPUT TEXT

قُلدُر سَرای هیبتِ، یامَن هُوَ الوِقار
تیغِ دو لَب شکار، مُسمی به ذِوالفَقار
#هوعلی

GROUND TRUTH

ترس (translation: fear)

PREDICTION

شادی (translation: joy)

MODEL CoT

متن داده شده به نظر میرسد که احساس شادی را نشان میدهد. از کلمات و عبارتهای استفاده شده، مانند 'قلدر سرای' و 'ذوالفقار'، نشان میآید که متن به یک شخصیت خیرپرداز و احترامآراست پرداخته است. این عبارتهای شعری به سمت شادی و تحسین یک فرد خاص منتهی میشوند.

Figure 11: An example of unfaithful CoT from deepseek-r1:32b (FA) on the PERSIANTEXTEMOTION dataset, categorized as a post-hoc reasoning.