When LLMs Annotate: Reliability Challenges in Low-Resource NLI

Solmaz Panahi[1], John Kelleher[2], Vasudevan Nedumpozhimana[2]

[1] Maynooth University, [2] Trinity College Dublin

Abstract

This paper systematically evaluates LLM reliability on the complex semantic task of Natural Language Inference (NLI) in Farsi, assessing six prominent models across eight prompt variations through a multi-dimensional framework that measures accuracy, prompt sensitivity, and intra-class consistency. Our results demonstrate that prompt design-particularly the order of premise and hypothesis-significantly impacts prediction stability. Proprietary models (Claude-Opus-4, GPT-4o) exhibit superior stability and accuracy compared to open-weight alternatives. Across all models, the 'Neutral' class emerges as the most challenging and least stable category. Crucially, we redefine model instability as a diagnostic tool for benchmark quality, demonstrating that observed disagreement often reflects valid challenges to ambiguous or erroneous gold-standard labels.

1 Introduction

LLMs have emerged as a potential solution to data annotation limitations. Several works, across a range of tasks, have harnessed the capabilities of powerful LLMs as teachers to generate and annotate dataset samples (Viswanathan et al.; He et al., 2024b; Ding et al., 2023; Gilardi et al., 2023; Chiang and yi Lee, 2023; Yadav et al., 2024; Wang et al., 2021). Applications include machine translation (de Gibert et al., 2025), classification (Li et al., 2024a), and open-ended tasks such as summarization and question answering (Honovich et al., 2022). These approaches capture the teacher model's semantic insights and reasoning capabilities, enabling knowledge transfer for training smaller or more efficient models.

While using LLMs as a proxy for human annotators has become increasingly popular, their reliability remains challenging. LLMs have limitations as direct human replacements, particularly in low-resource and nuanced task settings (Cheng et al., 2024). These limitations can be broadly categorized into methodological instabilities, inherent systematic biases, and task-specific constraints. LLMs responses are highly sensitive to minor changes in prompt phrasing, formatting, and exemplar selection (Sclar et al., 2024; Long et al., 2024). Studies have documented significant order effects, where the sequence of presenting information fundamentally alters model reasoning (Rezagholizadeh et al., 2025; Chen et al., 2024; Zhuo et al., 2024; Arakelyan et al., 2024), and have quantified their vulnerability to spurious features in prompt phrasing (Sclar et al., 2024). Apart from order, the format of prompts has been shown to have significant impact on LLMs performance (He et al., 2024a). This instability is further compounded in multilingual contexts, where prompt translation strategies (Tsarfaty et al., 2025) and language-specific templates (Gan and Mori, 2023) dramatically impact performance, raising particular concerns about their reliability as evaluators in low-resource languages (Fu and Liu, 2025). LLMs can inherit and amplify bias present in their training data which leads to systematic errors that are not random but are skewed against underrepresented perspectives (Suhara et al., 2024; Gallegos et al., 2024). For low-resource languages like Farsi, this is a critical issue; an LLM trained on skewed data may fail to understand cultural context and nuance. While cultural bias is a primary concern for low-resource languages, tasks such as NLI introduce additional dimensions of bias that threaten annotation validity. These include lexical bias, where models rely on surface-level word overlap; demographic bias, where social stereotypes influence logical judgments; and veridicality bias, which concerns the unresolved tension between strict textual inference and the use of common-sense world knowledge (Geiger et al., 2020). More generally the performance of LLMs as annotators is highly task-dependent and often fails on subjective or complex

linguistic tasks. While they may achieve strong results on fact-based or simple classification tasks, their performance deteriorates on tasks requiring pragmatic reasoning, world knowledge, or the interpretation of subjective intent (Mirzakhmedova et al., 2024; Haq et al., 2025; Calderon et al., 2025; Wang et al., 2023). Surveys and empirical studies consistently synthesize these axes, concluding that the reliability of LLMs as annotators or judges is profoundly task- and setup-dependent. This observation motivates the need for standardized protocols and benchmarks for LLM-as-annotator and LLM-as-judge, as well as careful implementation and awareness of their limitations (Tseng et al., 2024; Chiang and yi Lee, 2023; Cheng et al., 2025; Wang et al., 2025; Li et al., 2024b; Tan et al., 2024). The critical question, therefore, shifts from if LLMs can replace human annotators to under what specific conditions—which tasks, languages, and methodological setups—their strengths can be reliably leveraged while mitigating their profound weaknesses. The literature and research are highly focused on English, and fewer scholars work on non-English (Joshi et al., 2025; Whitehouse et al., 2023; Kaddour and Liu, 2024; Samuel et al., 2024). LLMs are demonstrably weaker in low-resource languages due to pre-training data imbalance, leading to poorer grasp of grammar, style, and reasoning. This creates a double bind: the communities most in need of scalable annotation solutions are those for whom the current technology is least reliable. To address this gap, we present a comprehensive reliability analysis of LLMs as annotators for Farsi NLI—a complex semantic task requiring nuanced logical reasoning that serves as a rigorous testbed for evaluating annotator capabilities in low-resource settings. Our contributions include:

- We introduce a comprehensive evaluation framework that goes beyond accuracy to assess LLM annotators. While individual metrics exist in prior work, their systematic integration for evaluating annotation reliability is novel. This methodology is language- and task-agnostic, providing a generalizable template for evaluating LLM annotator reliability across diverse linguistic contexts and annotation tasks beyond NLI.

- Empirical analysis of methodological instabilities, quantifying how prompt design induces prediction volatility.

- Evidence that benchmark quality substantially shapes perceived LLM reliability, with many high-instability cases aligning with genuine annotation ambiguity or noisy gold labels rather than clear model errors.

2 Methodology

Our methodology is designed to test several key hypotheses about LLM annotation behavior: that model performance varies systematically with prompt formulation, that annotation stability differs across model families, and that reliability metrics beyond accuracy reveal critical insights about model robustness. To test these hypotheses, we constructed a controlled experiment examining six LLMs across eight prompt conditions using ParsiNLU training split (750 samples: 275 entailment, 241 neutral, and 234 contradiction) as our evaluation testbed (Khashabi et al., 2021).

2.1 Annotator Models

LLMs differ substantially in scale, training data, and development methods. To systematically investigate how underlying size and training paradigms mediate annotator reliability, we selected different instruction-tuned LLMs based on access type (proprietary versus open-weight), model scale (parameter count and computational demands), and linguistic coverage (general-purpose versus explicitly multilingual). The models we selected are Claude-Opus-4, GPT-4o, GPT-4o-mini, Qwen3-30B-Instruct, Llama3.3-70B-Instruct, and Aya-23-35B. [1]

2.2 Prompt Design and Variations

The sensitivity of LLMs to prompt formulation is a well-documented phenomenon, fundamentally linked to their auto-regressive training objective. Since these models are trained to maximize sequence likelihood, their performance on a given task is governed by the probability they assign to both the prompt itself and the correct answer (Yao et al., 2023). This probabilistic dependency explains the observed instability in accuracy when prompts are altered. Our prompt design framework employs controlled variations across three

[1]We also evaluated smaller models in the 7B parameter range (including Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct) but excluded them from detailed analysis due to substantially degraded performance, with accuracy ranging between 53-56%.

critical dimensions: instruction language (English/-Farsi), stylistic elaboration (concise/detailed), and structural ordering (premise-hypothesis/hypothesis-premise), enabling precise isolation of each factor's influence on annotation stability. We design the system prompts and user prompts as follows:

System Prompts System prompts are used to define the role of the language model and constrain its behavior during the classification task. In our experiments we use two variants of system prompt ((a) elaborate or (b) concise):

> *a) You are a helpful assistant that evaluates whether a hypothesis can be inferred from a premise. Answer using only one of the following labels: Entailment, Contradiction, or Neutral. Do not explain.*
> *b) Classify the relationship between the following two statements as Entailment, Contradiction, or Neutral. Respond with one word only.*

System prompts are provided in either English or Farsi to examine the influence of instruction language on model performance. We do not translate the input samples (premises and hypotheses). Therefore, all models receive the original (Farsi) samples with both English and Farsi instructions. The inclusion of bilingual system prompts allows for comparative analysis of how instruction language affects the behavior of multilingual models.

User Prompts To evaluate the sensitivity of language models to input phrasing, two user message formulations ((a) original or (b) reverse) are employed:

> *a)Premise:{premise}Hypothesis:{hypothesis}. Answer:*
> *b)Hypothesis:{hypothesis}Premise:{premise}. Answer:*

Due to the directional nature of entailment, we employ explicit semantic role labels (Premise: and Hypothesis:) that persist across both ordering conditions. This design ensures that models receive clear, unambiguous identification of each component's logical role, thereby isolating order effects from potential role confusion. By combining the four system prompt variants (two in each language) with the two user prompt formats described (`original` and `reverse`), a total of eight distinct prompt configurations are generated for each model (Table 1).

Prompt	Lang	Style	Order
`original_elaborate_en`	En	Elaborate	original
`original_concise_en`	En	Concise	original
`original_elaborate_fa`	Fa	Elaborate	original
`original_concise_fa`	Fa	Concise	original
`reverse_elaborate_en`	En	Elaborate	reverse
`reverse_concise_en`	En	Concise	reverse
`reverse_elaborate_fa`	Fa	Elaborate	reverse
`reverse_concise_fa`	Fa	Concise	reverse

Table 1: Prompt variants used to evaluate LLM sensitivity to linguistic and structural variations. Each prompt combines language (English, Farsi), instruction style (elaborate, concise), and premise-hypothesis ordering (original, reverse).

2.3 Evaluation metrics

We evaluate LLM reliability through multiple robustness dimensions: accuracy measures task correctness; sensitivity quantifies prompt-induced decision instability by measuring normalized entropy of predictions across prompt variants (Errica et al., 2025); consistency assesses class-wise stability (Errica et al., 2025); and inter-model agreement captures consensus across model families. Model evaluation was conducted using a zero-shot prompting strategy for all models. Generation parameters were controlled by setting the sampling temperature to $T = 0$ and limiting the maximum output sequence length to 10 tokens.

3 Results

3.1 Accuracy

We compute the model's prediction accuracy independently for each prompt variants. This prompt-level evaluation enables a finer-grained diagnosis of prompt sensitivity: certain prompt variants may induce significantly higher or lower accuracy for the same model and input distribution. Figure 1 presents the accuracy distribution across prompts and models. For all models, accuracy was generally higher when the prompt order was `original` rather than `reversed`. Claude-Opus-4 achieved the highest accuracy (0.83) with the `original_concise_fa` prompt, while Aya-23-35B had the lowest accuracy (0.61) under the `reverse_elaborate_fa` prompt. Although all models reached their maximum accuracy with original order prompts, there was noticeable variation

in average performance and prompt sensitivity between models.

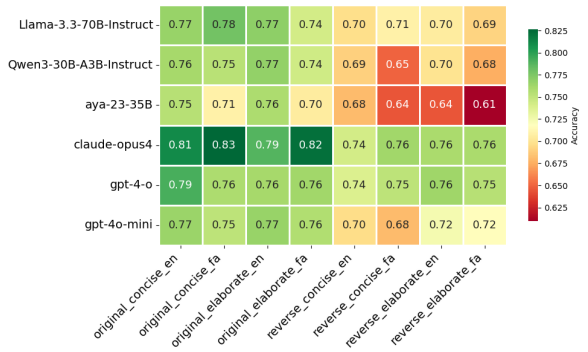

Figure 1: Accuracy of six language models across eight prompt configurations for Farsi NLI. Original premise-hypothesis ordering (left four columns) consistently yields higher accuracy than reversed ordering (right four columns).

3.2 Sensitivity

This metric captures the degree to which a model's predicted labels for a given input vary across semantically equivalent prompt formulations, independent of ground truth. Specifically, sensitivity is measured as the *normalized entropy* of the predicted label distribution over a set of prompt variants for the same input (Errica et al., 2025). Formally, for an input x and a set of prompt variants $\{p_1, \ldots, p_n\}$, let $P_x \in R^K$ denote the empirical distribution of predicted labels over K classes. The sensitivity for input x is then computed as:

$$\text{Sensitivity}(x) = \frac{H(P_x)}{\log_2 K} \quad (1)$$

Where $H(P_x) = -\sum_{i=1}^{K} P_x(i) \log_2 P_x(i)$. Sensitivity does not evaluate the correctness of a model's prediction, but rather its robustness to prompt variation– that is, the extent to which its outputs remain stable when presented with alternative phrasings of the same underlying instruction. Averaging this value across a set of inputs provides a model-level sensitivity score. This metric enables analysis beyond traditional accuracy-based evaluation by diagnosing *prompt-induced decision instability*.

Figure 2 illustrates the distribution of prompt sensitivity scores for all evaluated models. Each violin represents the distribution of sample-level sensitivity scores, where a lower score indicates consistent predictions across prompt variants, and a higher score denotes greater variability or instability. The proprietary models-Claude-Opus-4 and GPT-4o-show narrower distributions concentrated near zero, confirming their robustness to prompt phrasing and alignment stability. In contrast, open-weight models exhibit broader distributions, with substantial density at higher sensitivity scores, reflecting less robustness to prompt variants. Overall, the figure underscores that prompt robustness is not uniform across models, and that even state-of-the-art open-weight systems can exhibit substantial behavioral inconsistency under seemingly minor prompt variations. Samples with sensitivity values above 0.67 [2] typically represent cases where the model shows significant confusion between all three classes (entailment, contradiction, neutral), reflecting high uncertainty in decision boundaries.

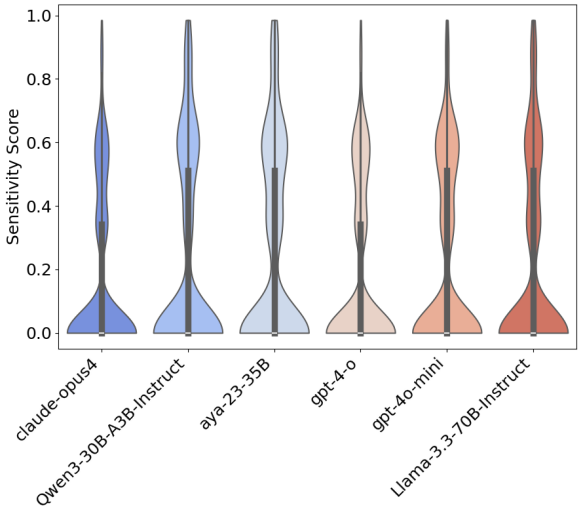

Figure 2: Violin plots of prompt sensitivity scores for six LLMs. Higher sensitivity scores indicate greater instability in responses. Proprietary models (Claude-Opus-4, GPT-4o) show lower median sensitivity and more compact distributions, reflecting consistent behavior across prompts. In contrast, open-weight models (LLaMA-3.3-70B, Aya-23-35B) display broader distributions and higher variability.

We next focus on the subset of zero-sensitivity samples—instances for which a model consistently predicts the same label across all prompt variants. These samples represent cases of complete intra-model agreement, suggesting high internal confidence and robustness to prompt perturbation. By analyzing how these stable predictions align with the gold-standard annotations, we evaluate whether prediction stability necessarily implies correctness.

[2]The normalized entropy threshold of 0.67 was determined through empirical analysis of vote distributions across eight prompt variations, representing the point at which model predictions transition from clear class preference ($\geq 75\%$votes for one class) to significant inter-class confusion.

Figure 3 shows that proprietary models, particularly Claude-Opus-4 and GPT-4o, exhibit both higher numbers of zero-sensitivity samples and a greater proportion of correct predictions, suggesting a stronger alignment between stability and semantic accuracy. In contrast, open-weight models such as Aya-23-35B and Qwen-3-30B demonstrate a larger share of consistently incorrect predictions, indicating that stability alone does not guarantee correctness. This pattern implies that certain models may internalize biases or spurious correlations that yield confident yet systematically flawed reasoning. Overall, the analysis highlights a key limitation of using prediction consistency as a proxy for annotation reliability—while stability reflects internal coherence, it must be interpreted in conjunction with gold-standard alignment to assess the true trustworthiness of LLM-based annotations

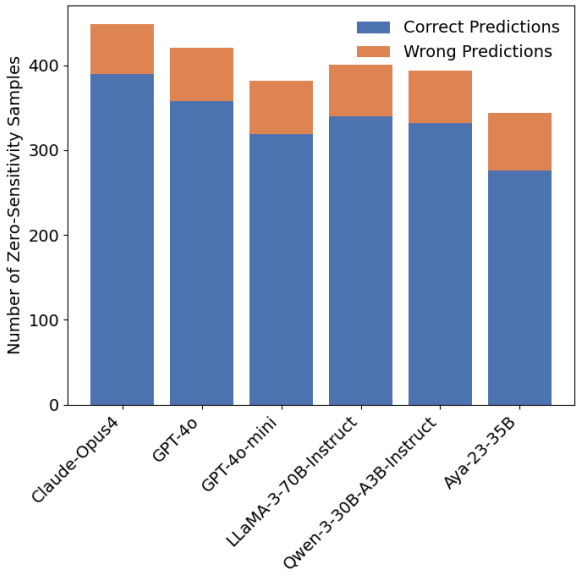

Figure 3: Correct and incorrect predictions among zero-sensitivity samples for each model (total dataset: 750 samples). Proprietary models achieve higher overall stability with fewer confident errors, whereas open-weight exhibit a greater proportion of consistently incorrect predictions.

3.3 Consistency (Class-Level)

This metric assesses whether a model generates similar predictive behavior across different inputs belonging to the same class. Unlike sensitivity, which captures per-input-prompt variability, consistency is defined over a set of inputs and reflects the alignment of predictions across those inputs under prompt variation (Errica et al., 2025). Formally, let $\mathcal{D}_c = \{x_1, x_2, \ldots, x_n\}$ be the subset of inputs labeled with class c in the gold dataset, and let $P_{x_i} \in R^K$ denote the empirical distribution over K output classes predicted for input $x_i \in D_c$ across multiple prompt variants. The *Total Variation Distance* (TVD) between two such distributions P_{x_i} and P_{x_j} is defined as:

$$\text{TVD}(P_{x_i}, P_{x_j}) = \frac{1}{2} \sum_{k=1}^{K} \left| P_{x_i}(k) - P_{x_j}(k) \right| \quad (2)$$

The *consistency* for class c is then calculated as:

$$\text{C} = 1 - \frac{2}{n(n-1)} \sum_{x_i, x_j \in D_c} \text{TVD}(P_{x_i}, P_{x_j}) \quad (3)$$

where $n = |\mathcal{D}_c|$ is the number of examples with gold label c. A higher consistency value (closer to 1) indicates that the model produces more uniform predictions across examples of the same class, while lower values suggest divergent or unstable intra-class behavior.

Ideally, a consistent LLM produces similar output distributions for inputs that share the same gold label, even under minor prompt rephrasings. Consistency therefore captures the extent to which a model generalizes its predictive behavior within a class across prompt variants. Importantly, consistency can remain high even when a model exhibits sensitivity to prompt variation; in such cases, high consistency indicates systematic rather than random errors, whereby similar inputs fail in similar ways across prompts. This structured behavior makes errors predictable and amenable to targeted prompt engineering, whereas low consistency reflects erratic intra-class behavior for which prompt interventions are less likely to be effective.

Table 2 reveals distinct patterns in how different models handle NLI task under prompt variation. First, for the majority of models the lowest consistency score is on the Neutral class, indicating that many models are inconsistent in their predictions across prompt variations. Second, the models exhibit significant specialization, with different models excelling in different classes; no single model leads in all three categories. Claude-Opus-4 and GPT-4o demonstrate the most balanced performance across all classes, while others show a pronounced strength in one or two classes at the expense of the third.

Model	Entailment	Contradiction	Neutral
Aya-23-35B	0.83	0.62	0.48
Qwen3-30B-A3B	0.76	0.71	0.45
LLaMA-3.3-70B	0.79	0.69	0.48
Claude-Opus-4	0.62	0.72	0.67
GPT-4o-mini	0.62	0.79	0.47
GPT-4o	0.56	0.78	0.64

Table 2: Class-wise consistency scores (1 - TVD) across models and NLI classes. Higher values indicate more uniform within-class predictions. While all models show the lowest consistency for Neutral, Claude-Opus-4 and GPT-4o demonstrate the most balanced performance across all three classes, whereas open-weight models show pronounced specialization.

3.4 Inter-Model Agreement: Cohens Kappa and Fleiss' Kappa

Beyond individual model behavior, we additionally examine inter-model agreement patterns to assess whether different LLMs converge on consistent annotations or exhibit systematic differences in their interpretive frameworks. To quantify consensus across different LLMs and prompt configurations, we used two established statistical measures of inter-annotator agreement. Cohen's Kappa assesses pairwise agreement between two raters, while Fleiss' Kappa generalizes this framework to multiple raters, both accounting for agreement expected by chance. In our experimental context, these metrics transform our experimental setup into a multi-annotator study, where each model-prompt combination functions as an independent rater. This approach allows us to measure whether diverse LLMs converge toward a consistent interpretive framework for NLI or exhibit fundamental divergences in their reasoning. Figure 4 shows pairwise Cohen's Kappa scores across all prompt variants, revealing significant variation in alignment between models, from strong agreement (e.g., Claude-GPT-4o at 0.79) to weak consensus (e.g., Aya-23-35B with other models).

Table 3 shows the Fleiss' Kappa scores across different models. To contextualize our model agreement scores, we refer to the inter-annotator agreement reported in the ParsiNLU annotation procedure (Khashabi et al., 2021), where three human annotators achieved a Fleiss' Kappa of 0.77 on a randomly selected subset of 150 examples. Proprietary models consistently achieve higher Fleiss' Kappa scores across most prompt conditions, with particularly strong consensus (0.73). In contrast, the open-weight model's best agreement is 0.69. The overall lower agreement when combining all

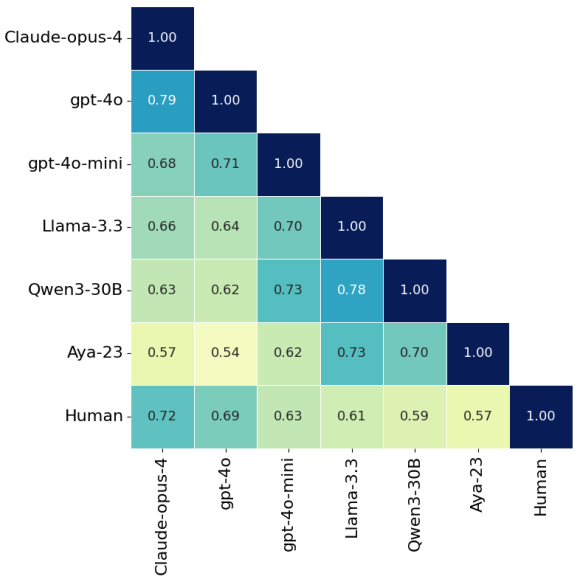

Figure 4: Pairwise Cohen's Kappa scores across all prompt variants, indicating the degree of agreement between model predictions. Claude-Opus4 and GPT-4o exhibit the highest agreement, while lower agreement is observed between Aya-23 and the other models.

models underscores fundamental behavioral differences between architectural families, suggesting that model diversity introduces substantial variation in NLI interpretation strategies. These findings indicate that both model diversity and prompt design play critical roles in inter-model consistency, highlighting the value of prompt sensitivity analysis when using ensembles or comparing model outputs.

Model Group	reverse_elaborate_en	reverse_concise_en	original_elaborate_en	original_concise_en	reverse_elaborate_fa	reverse_concise_fa	original_elaborate_fa	original_concise_fa
Open-Models	0.65	0.63	0.69	0.69	0.61	0.59	0.68	0.69
Close-Models	0.65	0.68	0.72	0.73	0.69	0.65	0.70	0.70
All-Models	0.62	0.62	0.67	0.69	0.55	0.56	0.61	0.65

Table 3: Inter-model agreement (Fleiss' Kappa) across prompt variants by model type. The overall lower agreement for All-models reflects the behavioral diversity between proprietary and open-weight systems.

3.5 Probing the Instabilities

A central question in evaluating LLMs as annotators is whether observed instability reflects fundamental model limitations or inherent challenges in the annotation task and data quality itself. While NLI appears to demand objective logical reason-

ing, human language interpretation is inherently influenced by context, pragmatics, and real-world knowledge. This creates tension between *semantic* and *pragmatic* inference. Given that, we identify the subset of examples exhibiting the highest prompt sensitivity. High-sensitivity cases are defined as inputs for which a model's predicted label changes frequently across prompt variation. We focused on instances where all evaluated models demonstrated high sensitivity, suggesting inherent ambiguity in the examples themselves. To understand the decision mechanisms underlying label predictions in high-sensitivity cases, we implemented chain-of-thought prompting to make models' internal reasoning processes explicit. We replicated the prompt variation experiments on this subset, modifying the system prompts to include `"Explain your reasoning"` instructions while maintaining the same structural and linguistic variations. In the right column, we provide an example of these subjective cases, illustrating how the same model—when prompted differently—articulates divergent yet logically sound interpretations of the same premise-hypothesis pair, reflecting the inherent subjectivity that also challenges human annotation. This variation suggests that high-sensitivity samples often reflect inherent ambiguity in the task rather than random model error.

Finally, building on the earlier sensitivity analyses, we conducted a manual review of the subset of instances in which the Claude-Opus-4 (the state-of-the-art model in our experiments) exhibited non-zero sensitivity, resulting in a subset of 198 samples. To assess annotation reliability, we compared the predictions on these samples with the human-annotated gold labels. Out of the total, 97 cases showed disagreement between the model and the gold annotations. Upon inspection, the majority of these discrepancies were found to stem from errors or weaknesses in the gold dataset, rather than from model misclassification. Many instances were either ambiguous, excessively short, or lacking sufficient semantic content to support an unambiguous inference decision. Such noisy samples are removed or relabeled to create a clean version of the dataset. Following this data cleaning procedure, we re-evaluated the accuracy of the models using the refined dataset. By cleaning data, we ensured that the evaluation was conducted on a more reliable ground truth. Figure 5 shows the corresponding results after noisy and ambiguous samples were removed. The comparison between Figure 1 and Figure 5 reveals a consistent improvement across all evaluated models, with gains of up to several percentage points depending on the prompt variant.

Gold data sample

Premise: For many years, these things no longer make me happy. **Hypothesis**: Celebrations and joy are always good and pleasant, but now I am at an age where very few things make me happy.
Gold Label: Entailment.

GPT-4o-mini Reasoning

Prompt: `original_elaborate_en`
CoT Reasoning: *Entailment*. The premise states that "for many years, these things no longer make me happy," which implies a change in emotional response to things that used to bring joy. The hypothesis suggests that celebrations and joy are generally good and pleasant, but the speaker is now at an age where very few things make them happy. This aligns with the premise, as it indicates a decrease in happiness over time, supporting the idea presented in the hypothesis.
Prompt: `original_concise_en`
CoT Reasoning: *Neutral*. The premise states that the speaker has not been happy about certain things for a long time, indicating a lack of joy or satisfaction. The hypothesis suggests that celebrations and joy are generally good and pleasant, but the speaker is now at an age where very few things make them happy. While both statements touch on the theme of happiness, the premise does not directly support or contradict the hypothesis. The hypothesis acknowledges that joy exists but implies that the speaker's personal experience has changed, which aligns with the premise but does not necessarily follow from it.
Prompt: `reverse_concise_en`
CoT Reasoning: *Contradiction*. The hypothesis states that celebrations and joy are always good and pleasant, but the speaker is currently at an age where very few things make them happy. The premise supports this by stating that for a long time, these things no longer make the speaker happy. This creates a contradiction because if celebrations and joy are always good and pleasant, then it would imply that they should still bring happiness, which is not the case according to the premise. Therefore, the two statements cannot both be true at the same time.

The improvement is particularly pronounced for closed-source models such as Claude-Opus4 and GPT-4o, which already demonstrated higher baseline performance. These models benefit more strongly from the removal of mislabeled and ambiguous cases, suggesting that their predictions were often correct but had been penalized by unreliable ground-truth annotations. By contrast, open-weight models show smaller relative improvements. The performance gap between original and cleaned

datasets should be interpreted as revealing which model types are more affected by ambiguous annotations, rather than as validation of absolute model reliability.

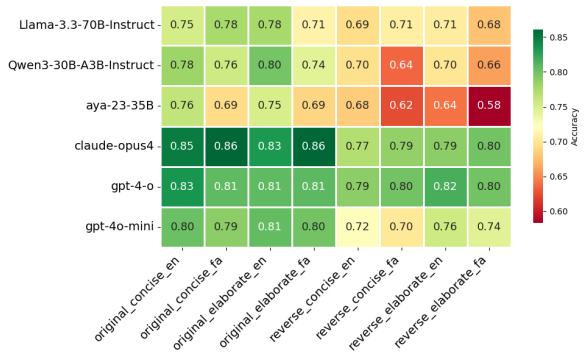

Figure 5: Accuracy comparison of six language models across eight prompt configurations on the refined (cleaned) datasets. The consistent performance improvement on the cleaned data suggests that a portion of the observed performance drop on the original benchmark is attributable to problematic or noisy gold labels, rather than inherent model weakness.

4 Discussion

This study set out to answer a central question: *How reliable are large language models when used as automatic annotators in low-resource settings.* The key insights, which we elaborate below, demonstrate that while LLMs hold significant promise, their dependable deployment requires a nuanced understanding of their systematic failure modes and behavioral instabilities.

Higher Reliability in Proprietary Models Our results demonstrate a clear reliability hierarchy between model families. Proprietary models (Claude-Opus-4, GPT-4o) exhibited significantly greater robustness, with lower sensitivity scores and better alignment between prediction stability and correctness (Figure 3). This suggests that their lower sensitivity scores and better stability-accuracy alignment reflect fundamentally more reliable reasoning patterns rather than mere scale advantages. For practical deployment, this creates a cost-reliability trade-off where proprietary models offer more reliable annotations despite their opacity and cost, while open-weight models offer viable but noisier alternatives.

Prompt Design as a Critical Reliability Factor The substantial impact of prompt formulation on annotation quality directly addresses the reproducibility concerns in using LLMs as annotators. Most notably, we found that ordering (premise-hypothesis sequence) induced significantly greater prediction volatility than language choice or instruction style in terms of sensitivity. We hypothesize that the reduced labeling performance under reverse ordering is rooted in model training practices, particularly the consistent exposure to sequences where the premise precedes the hypothesis.
However, language choice still influences overall accuracy-as indicated in our grouped analyses by prompt language and style (Figure 1), suggesting that while structural consistency supports reliability, linguistic formulation continues to play a role in final performance.

The Neutral Class as a Reliability Frontier Across all models and metrics, the Neutral class emerged as the primary reliability challenge, exhibiting the lowest consistency scores (Table 2). This pattern suggests that current LLMs struggle with the pragmatic reasoning required for partial entailment detection as well as with the inherent ambiguity of the Neutral category itself. The low consistency suggests that models struggle with the contextual inferences and world knowledge required for partial entailment judgments. This implies that Neutral classifications should be treated as lower-confidence predictions requiring additional verification in annotation workflows.

Benchmark Quality as a Hidden Reliability Variable Our analysis revealed that approximately 49% of high-sensitivity cases involved questionable gold labels or inherent ambiguities. After cleaning the dataset, we observed consistent performance improvements across all models (Figure 5), with proprietary models benefiting most significantly. This suggests that what appears as model instability may sometimes reflect legitimate disagreement with problematic annotations. Rather than treating this refined dataset as a purer ground truth, we view it as a diagnostic subset that highlights the impact of benchmark uncertainty on evaluation. The observed improvements should therefore be interpreted as upper bounds under reduced annotation noise, not as absolute indicators of model competence. However, analysis of models reasoning on high-sensitive samples reveals that LLMs are not merely relying on surface-level heuristics but are capable of accessing different, and often equally valid, reasoning paths. Our chain-of-

thought analysis demonstrates that the same model, when provided with different prompts, can articulate distinct yet logically sound interpretations for the same premise-hypothesis pair. This variability is not inherently a sign of model failure, but rather an indication that the NLI task itself contains genuine subjectivity-mirroring the interpretive disagreements observed among human annotators. The fact that models can be steered towards these different valid interpretations through prompt engineering suggests that their "instability" in these cases is better understood as contextual reasoning flexibility. Overall, these findings highlight the importance of rigorous dataset curation in benchmarking LLMs. This concern is compounded by prior work (Jiang et al., 2023; Plank, 2022), where human label variation arises from legitimate differences in interpretation, suggesting that some gold labels may not fully represent the spectrum of valid judgments. This observation calls for a more nuanced approach to evaluating model performance, particularly in low-resource settings where benchmark quality and consistency have an outsized impact on evaluation.

5 Conclusion

In this paper, we presented a comprehensive analysis of the reliability of LLMs as annotators for the NLI task in Farsi. The findings underscore the critical dependence of annotation reliability on prompt formulation, model families, and linguistic nuances. Collectively, these interpretations demonstrate that LLM reliability is not merely a function of model capability but emerges from the complex interaction between model development, task formulation, and benchmark quality-necessitating holistic approaches to deploying LLMs as annotators in low-resource contexts. In future work we will explore the effect of LLM annotators on the downstream task.

6 Limitations

While this study offers a detailed investigation into the reliability of LLMs as annotators for Farsi NLI, several limitations should be acknowledged. First, this study is limited to Farsi and the ParsiNLU dataset, which restricts the generalizability of findings to other low-resource languages, tasks, or annotation domains. Second, while our prompt variations are systematic, they do not encompass all possible prompt formulations—such as few-shot examples, chain-of-thought variations, or domain-specific instructions—which could further affect model stability and reliability. Third, we cannot rule out data contamination, as ParsiNLU has been public since 2021. While observed prompt sensitivity suggests genuine reasoning rather than memorization, partial exposure could inflate absolute performance metrics. Finally, the analysis does not assess the downstream impact of using LLM-generated labels for training or evaluating other models, leaving open questions about how annotation-level instabilities propagate in practical applications.

7 Acknowledgement

This publication has emanated from research supported in part by a grant from Taighde Éireann – Research Ireland under Grant number 18/CRT/6049. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of nli models. *Preprint*, arXiv:2401.14440.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *Preprint*, arXiv:2501.10970.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *ArXiv*, abs/2402.08939.

Lu Cheng, Mansooreh Karami, Huan Liu, Jundong Li, Song Wang, Amrita Bhattacharjee, Dawei Li, Zhen Tan, Alimohammad Beigi, and Bohan Jiang. 2024. Large language models for data annotation: A survey. *Preprint*, arXiv:2402.13446.

Lu Cheng, Kai Shu, Tianhao Wu, Huan Liu, Canyu Chen, Amrita Bhattacharjee, Yuxuan Jiang, Dawei Li, Zhen Tan, Alimohammad Beigi, Bohan Jiang, Chengshuai Zhao, and Liangjie Huang. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource

mt via synthetic data generation with llms. *Preprint*, arXiv:2505.14423.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2025. What did i do wrong? quantifying llms' sensitivity and consistency to prompt engineering. *Preprint*, arXiv:2406.12334.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual llm-as-a-judge? *Preprint*, arXiv:2505.12201.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Preprint*, arXiv:2309.00770.

Chengguang Gan and Tatsunori Mori. 2023. Sensitivity and robustness of large language models to prompt template in japanese text classification tasks. *Preprint*, arXiv:2305.08714.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Pushing the limits of rule-based reasoning in transformers through natural language satisfiability. *arXiv preprint arXiv:2012.10052*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Muhammad Uzair Ul Haq, Davide Rigoni, and Alessandro Sperduti. 2025. Llms as data annotators: How close are we to human performance. *Preprint*, arXiv:2504.15022.

Jia He, Arshdeep Sekhon, David Koleczek, Mukund Rungta, Franklin X Wang, and Sadid Hasan. 2024a. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024b. Annollm: Making large language models to be better crowdsourced annotators. *Preprint*, arXiv:2303.16854.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *Preprint*, arXiv:2212.09689.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Understanding and predicting human label variation in natural language inference through explanation. *Preprint*, arXiv:2304.12443.

Raviraj Joshi, Abhay Shanbhag, Suramya Jadhav, Amogh Thakurdesai, and Ridhima Sinare. 2025. On limitations of llm as annotator for low resource languages. *Preprint*, arXiv:2411.17637.

Jean Kaddour and Qi Liu. 2024. Synthetic data generation in low-resource settings via fine-tuning of large language models. *Preprint*, arXiv:2310.01119.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021. Parsinlu: A suite of language understanding challenges for persian. *Preprint*, arXiv:2012.06154.

Yinheng Li, Rogerio Bonatti, Sara Abdali, Justin Wagle, and Kazuhito Koishida. 2024a. Data generation using large language models for text classification: An empirical case study. *Preprint*, arXiv:2407.12813.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024b. Leveraging large language models for nlg evaluation: Advances and challenges. *Preprint*, arXiv:2401.07103.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? *CoRR*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mehdi Rezagholizadeh, Tanya Roosta, Peyman Passban, and Bryan Guan. 2025. The order effect: Investigating prompt sensitivity in closed-source llms. *Preprint*, arXiv:2502.04134.

Vinay Samuel, Houda Aynaou, Arijit Ghosh Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. Can llms augment low-resource reading comprehension datasets? opportunities and challenges. *Preprint*, arXiv:2309.12426.

Mikayla Sclar, Yejin Choi, Yulia Tsvetkov, and Maarten Sap. 2024. Quantifying language models' sensitivity to spurious features in prompt design. *arXiv preprint arXiv:2402.16892*.

Yoshi Suhara, Rickard Stureborg, and Dimitris Alikaniotis. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *Preprint*, arXiv:2402.13446.

Reut Tsarfaty, Tzuf Paz-Argaman, and Itai Mondshine. 2025. Beyond english: The impact of prompt translation strategies across languages and tasks in multilingual llms. *Preprint*, arXiv:2502.09331.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are expert-level language models expert-level annotators? *Preprint*, arXiv:2410.03254.

Vijay Viswanathan, Xiang Yue, Alisa Liu, Yizhong Wang, and Graham Neubig. Synthetic data in the era of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Saizhuo Wang, Jian Guo, Wei Li, Wen Gao, Kun Zhang, Shengjie Ma, Chengjin Xu, Xuhui Jiang, Yinghan Shen, Yuanzhuo Wang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Jiawei Gu, Honghao Liu, and Lionel Ni. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *Preprint*, arXiv:2108.13487.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Sachin Yadav, Tejaswi Choppa, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4. *Preprint*, arXiv:2407.04130.

Shunyu Yao, Thomas L. Griffiths, Dan Friedman, R. Thomas McCoy, and Matthew Hardy. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *Preprint*, arXiv:2309.13638.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. *Preprint*, arXiv:2410.12405.