



Bridging the Data Gap: LLM-Driven Synthetic Data Generation for Low-Resource Natural Language Inference

Solmaz Panahi

Supervisors:

David Malone
John Kelleher
Vasudevan Nedumpozhi

*A thesis submitted in fulfillment of the requirements
for the Ph.D. degree in Data Science*

at the

Hamilton Institute
Maynooth University
Maynooth, Co. Kildare, Ireland

August 31, 2026

To the people of Iran

Summary

This thesis addresses the challenge of data scarcity in low-resource languages by leveraging the dual role of Large Language Models (LLMs) to both generate and annotate data. We evaluate the reliability of LLM-based annotations and examine the effectiveness of LLM-generated synthetic data in improving downstream task performance. This research uses the Natural Language Inference (NLI) task in Farsi as its case study. Our approach involves benchmarking both open-weight and closed-source LLMs on a gold-standard dataset, developing an LLM-driven pipeline to generate labeled premise-hypothesis pairs, and training mT5 models to measure performance gains. Further experiments were conducted with various LLM annotators to analyze how annotation quality affects model outcomes.

The results reveal key insights. First, synthetic data substantially improves model performance, particularly under sequential training regimes, confirming its effectiveness as a strategy for data augmentation in low-resource settings. Second, the quality of annotations plays a decisive role in downstream outcomes—higher-quality labels consistently lead to greater accuracy and robustness. Finally, the reliability of an LLM as an annotator emerges as a multifaceted property, shaped by factors including model architecture, prompt design, and the inherent complexity of the task itself.

These findings show that while LLM-generated synthetic data can mitigate data scarcity issues, annotation quality remains a critical factor influencing model performance. The thesis contributes insights into optimizing annotation pipelines and synthetic data utility for low-resource NLP, with implications for multilingual model training strategies.

Declaration

I, Solmaz Panahi, declare that this thesis titled, “**Bridging the Data Gap: LLM-Driven Synthetic Data Generation for Low-Resource Natural Language Inference**” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this university.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Solmaz Panahi

31-08-2026

Acknowledgements

The completion of this thesis would not have been possible without the support and guidance of many individuals to whom I owe my deepest gratitude.

First and foremost, I wish to express my sincere appreciation to my supervisors for their invaluable guidance, unwavering patience, and insightful feedback throughout this research journey. Their expertise and encouragement were instrumental in shaping this work.

My heartfelt thanks go to my family for their unconditional love, endless patience, and steadfast belief in me. Their emotional support was my anchor during the most challenging phases of this work.

I am fortunate to have been surrounded by wonderful friends and colleagues who made this journey less daunting. I extend my gratitude to my officemates at Hamilton Institute for fostering a friendly environment.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research Questions	3
1.3	Thesis Overview	4
1.4	Contributions	5
2	Background and Related Work	7
2.1	The Linguistic Divide in NLP: From Multilingual Models to General-Purpose LLMs	7
2.2	From Human to LLM-based Annotation	8
2.3	Challenges and Limitations of LLM-based Data Synthesis	10
2.4	Farsi NLP	12
2.4.1	Linguistic Challenges	12
2.4.2	Farsi as a Representative of Low-Resource Languages	14
2.4.3	State of the Art Across Farsi NLP Tasks	15
2.5	Natural Language Inference	16
2.5.1	Benchmarks and Their Limitations	17
2.5.2	Subjectivity in NLI Annotation	17
2.5.3	Case Study: ParsiNLU	19
2.6	Selected Large Language Models	20
2.7	Prompt Engineering	21
3	LLMs as Annotators for NLI	24
3.1	Introduction	24
3.2	Method	24
3.2.1	Prompt Design and Variations	25
3.2.2	Post-processing and Label Normalization	26
3.3	Results	27
3.3.1	Accuracy	27
3.3.2	Prompt Sensitivity	30

3.3.3	Consistency (Class-Level)	38
3.3.4	Probing the Instabilities	41
3.3.5	Inter-Model Agreement: Cohen’s Kappa and Fleiss’ Kappa	47
3.4	Discussion	50
3.5	Summary	53
4	LLM-Driven Data Generation and Labeling at Scale	54
4.1	Introduction	54
4.2	Method	55
4.3	Case Study: Data Generation for Farsi NLI	56
4.3.1	Premise and Hypothesis Generation	57
4.3.2	Labeled Data Generation	59
4.4	Lexical Analysis of Synthetic and Gold Data	61
4.4.1	Human Evaluation on Synthetic Data Quality	64
4.5	Experimental Design	69
4.6	Results	72
4.6.1	Synthetic data size effect	72
4.6.2	Training regime effect	73
4.6.3	Effect of Balancing Label Distribution	80
4.7	Discussion	81
4.8	Summary	82
5	The Role of Label Quality	83
5.1	Introduction	83
5.2	Method	84
5.2.1	Model Effect: General-Purpose vs Task-Specific Annotators	84
5.2.2	Prompts Effect: Within-Model Impact on Downstream Performance	85
5.2.3	Ensemble Effect: Quality Optimization Strategies	85
5.3	Results	86
5.3.1	Annotator Bias: Label Distribution Shifts Across Models and Prompts	86
5.3.2	Prompt-Variant Insights Across Models	90
5.3.3	Downstream Impact of Annotation Choices	91
5.4	Discussion	101
5.4.1	The Dominant Effect of Annotator Quality on Downstream Performance	101
5.4.2	Architectural Considerations: Specialization vs. Scale	101
5.4.3	The Amplifying Role of Prompt Design	102
5.4.4	The Promise and Cost of Ensemble Methods	102

CONTENTS

5.5 Summary	103
6 Conclusion	104
6.1 Summary of Thesis	104
6.2 Answering Research Questions	104
6.3 Limitations	105
6.4 Future Work	106
6.5 Concluding Remarks	107
Bibliography	108

List of Figures

2.1	Tokenization comparison of semantically equivalent English and Farsi text. The English sentence consists of 16 tokens, while the Farsi (right-to-left) sentence yields 29 tokens, highlighting differences in script and tokenization behavior across languages.	13
2.2	Distribution of natural and MNLI-derived samples across NLI labels in the ParsiNLU [66] dataset	20
3.1	Accuracy comparison of six language models across eight prompt configurations for the NLI task. Models are evaluated under combinations of language (English/Farsi), system prompt style (detailed/concise), and premise-hypothesis ordering. The highest recorded accuracy (0.83) was achieved by Claude-Opus-4 under the <code>original_concise_fa</code> prompt, while the lowest (0.61) was observed for Aya-23-35B under the <code>reverse_elaborate_fa</code> prompt. This visualization highlights prompt and model-dependent variability in performance for the Farsi NLI annotation task.	28
3.2	Per-Class accuracy comparison of six language models across eight prompt configurations for NLI tasks. Models are evaluated under combinations of language (English/Farsi), system prompt style (detailed/concise), and premise-hypothesis ordering. The results highlight strong and consistent performance on Entailment, greater variability and moderate accuracy for Contradiction, and substantially lower and more dispersed accuracy for Neutral, indicating that identifying Neutral relationships presents the greatest challenge for current LLMs in this setting.	29

3.3	Distribution of prompt sensitivity scores across six instruction-tuned LLMs. The violin plots illustrate how each model’s predictions vary under different prompt formulations. 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.	33
3.4	Comparison of model sensitivity distributions under different prompt orders. The upper plot shows sensitivity scores when the premise precedes the hypothesis (original order), while the lower plot reverses the order. Across all models, reversing the input sequence increases overall sensitivity, particularly for open-weight models such as LLaMA-3.3-70B and Aya-23-35B, indicating that the order of the premise and hypothesis substantially influences prediction stability.	35
3.5	Comparison of model sensitivity distributions under different system prompt languages (English vs. Farsi). The upper plot shows sensitivity scores for the English system prompt, and the lower plot shows sensitivity scores for the Farsi system prompt.	36
3.6	Comparison of model sensitivity distributions under different system prompt styles. The upper plot shows sensitivity scores when the system prompt is elaborated, and the lower plot shows sensitivities when the system prompt is concise.	37
3.7	Correct and incorrect predictions among zero-sensitivity samples for each model (total dataset: 750 samples). Each bar represents the total number of samples for which a model produced identical labels across all prompt variants (i.e., zero-sensitivity cases). The blue segment indicates predictions that align with the gold-standard labels, while the orange segment denotes consistent but incorrect predictions. Proprietary models such as Claude-Opus-4 and GPT-4o achieve higher overall stability with fewer confident errors, whereas open-weight multilingual models like Aya-23-35B and Qwen3-30B exhibit a greater proportion of consistently incorrect predictions, suggesting that stability does not always imply correctness.	38

3.8	Two-Class Confusion frequencies across models (total dataset: 750 samples). This chart displays the count of instances where model predictions were confused between pairs of classes (Entailment-Neutral, Contradiction-Entailment, Contradiction-Neutral) across different prompt variations. The results confirm that Entailment-Neutral is the most prevalent confusion, indicating a shared challenge in distinguishing these semantically subtle classes.	41
3.9	Accuracy comparison of six language models across eight prompt configurations for NLI tasks after refining the gold dataset.	47
3.10	Heatmap of pairwise Cohen’s kappa between all model–prompt configurations	51
4.1	Synthetic data generation pipeline for classification tasks with two textual inputs. An LLM, guided by task instructions, generates input pairs (x_1, x_2) from raw text. A second LLM, prompted with judgment instructions, labels the pairs, producing annotated triplets (x_1, x_2, y) for training.	56
4.2	Sample premise–hypothesis pairs generated by the synthetic data pipeline in Farsi, with English translations provided for clarity. The examples illustrate how the summarization-based approach naturally produces Entailment, Contradiction, and Neutral relationships.	58
4.3	Distribution of sentence lengths (in number of words) across gold-standard dataset splits. On average, premise sentences are approximately 23 words long, while hypotheses average around 11 words.	59
4.4	Distribution of sentence lengths for premise (left) and hypothesis (right) in the synthetic and gold training datasets. On average, synthetic premise sentences are approximately 33 words long, while synthetic hypotheses average around 19 words.	59
4.5	Label distribution (%) across different synthetic dataset sizes. The distribution remains skewed toward Entailment as dataset size increases, reflecting a consistent bias in the LLM-based summarization and labeling pipeline.	61
4.6	Lexical overlap ratios between premise and hypothesis across data sources. Synthetic data shows higher overlap, especially for Entailment, indicating stronger lexical similarity induced by summarization.	63
4.7	Premise length distributions across synthetic and gold datasets, grouped by label. Synthetic premises are generally longer, particularly in the Entailment class.	63

4.8	Hypothesis length distributions for synthetic and gold data. Synthetic hypotheses are longer overall, consistent with their origin as LLM-generated summaries of the premise.	64
4.9	Confusion matrix comparing GPT-4o-mini predictions with human-annotated labels on a synthetic dataset of 600 samples.	69
4.10	Zero-shot classification performance of GPT-4o-mini on the ParsiNLU test set.	72
4.11	Performance of mT5-small and mT5-base models across increasing synthetic data sizes (5k–50k samples) under three training configurations: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue)	73
4.12	Entailment class performance for mT5-small (top row) and mT5-base (bottom row) models across different synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded regions denote 95% confidence intervals over five random seeds. Precision, recall, and F1 consistently improve with larger data, with the Synthetic → Gold regime yielding the highest and most stable performance, particularly in recall and overall F1.	78
4.13	Contradiction class performance for mT5-small (top row) and mT5-base (bottom row) models across increasing synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded regions denote 95% confidence intervals over five random seeds. Precision and F1 improve steadily with larger data, while recall shows higher variability across seeds. The Synthetic → Gold regime achieves the most consistent and overall strongest performance, particularly for the mT5-base model.	79
4.14	Neutral class performance for mT5-small (top row) and mT5-base (bottom row) models across increasing synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded areas represent 95% confidence intervals across five random seeds. Precision and F1 steadily improve with data size for both models, while recall shows higher variance. The Synthetic → Gold training regime consistently achieves the highest and most stable performance across metrics, indicating better generalization and calibration for the Neutral class.	80
4.15	Performance of mT5-small (top row) and mT5-base (bottom row) trained on imbalanced and balanced synthetic datasets across increasing data sizes.	81

5.1	Comparison of pairwise confusion matrices between GPT-4o-mini and other annotators (Aya-worst, Aya-best, and DeBERTa-NLI) across NLI classes. GPT-4o-mini and Aya-best show the strongest alignment, especially for <i>Entailment</i> and <i>Neutral</i> . Aya-worst and DeBERTa-NLI display larger off-diagonal errors, indicating greater divergence in annotation behavior and decision boundaries.	88
5.2	Comparison of pairwise confusion matrices between GPT-4o-mini and other annotators (Prompts-ensemble and Models-ensemble) across NLI classes. GPT-4o-mini and Models-ensemble exhibit the strongest alignment, especially for <i>Entailment</i> and <i>Neutral</i> . Prompts-ensemble is highly skewed toward <i>Entailment</i>	89
5.3	Correlation between prompt-level agreement with majority-vote labels and gold-standard accuracy across eight prompt variants for each model. Strong positive correlations for Claude-Opus-4 ($r=0.96$), Aya-23-35B ($r=0.87$), and Llama3-70B ($r=0.80$) indicate that ensemble agreement serves as a reliable proxy for annotation quality. In contrast, weaker correlations for GPT variants ($r=0.51-0.55$) and Qwen3-30B ($r=0.46$) reflect greater prompt-induced performance variability. Least-squares regression lines and Pearson correlation coefficients demonstrate the strength of the relationship for each model family.	91
5.4	Performance of mT5-small and mT5-base under <i>Synthetic</i> training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. For both model sizes, random or low-quality annotations (Aya-worst) yield substantially lower scores, confirming the importance of annotation quality.	93
5.5	Performance of mT5-small and mT5-base under <i>Synthetic(Ours)+Gold</i> training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. In the mT5-small model, the Aya-best annotator consistently outperforms all other conditions. In contrast, for mT5-base, the Models-ensemble and DeBERTa-NLI annotators achieve performance comparable to Aya-best.	93
5.6	Performance of mT5-small and mT5-base under <i>Synthetic(Ours)→Gold</i> training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. Under synthetic $\rightarrow$ gold fine-tuning, mT5-small treats prompts-ensemble and model-ensemble annotators near-identically, while mT5-base shows a clear preference for the prompts-ensemble over the model-ensemble.	94

LIST OF FIGURES

5.7	Per-Class Performance of mT5-small (Synthetic $\rightarrow$ Gold) across annotator quality conditions.	98
5.8	Per-Class Performance of mT5-base (Synthetic $\rightarrow$ Gold) across annotator quality conditions.	99

List of Tables

2.1	Overview of Farsi NLP datasets. "Type" indicates whether data is natively authored or derived via translation/crawling.	15
2.2	Examples of premise-hypothesis pairs and their corresponding NLI labels.	16
2.3	Label distribution of the ParsiNLU [66] dataset across train, validation, and test splits.	20
2.4	Overview of LLMs evaluated in this study, including model scale, access type, and multilingual benchmark performance (MMMLU).	21
3.1	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).	26
3.2	Multi-dimensional evaluation framework for assessing LLM annotator reliability. The taxonomy distinguishes between prompt-level, input-level, class-level, and model-level analyses to provide a holistic view of annotation stability.	27
3.3	Mean sensitivity scores across all prompt variants for each model, disaggregated by inference class (Entailment, Contradiction, Neutral). Lower sensitivity values indicate greater robustness to prompt reformulation, whereas higher values reflect greater variability in predictions across different prompt phrasings. Among the evaluated models, Claude-Opus-4 and GPT-4o exhibit the greatest stability, while Aya-23-35B shows the greatest prompt sensitivity—particularly for the Neutral class.	32
3.4	Mean sensitivity across models by factor—language (EN vs. FA), input order (Original vs. reverse), and prompt style (Elaborate vs. Concise). Higher values indicate greater prediction volatility; sensitivities are averaged over four prompt variants within each experimental condition (prompt language and input ordering).	34

3.5	Class-wise consistency scores (1 - TVD) for each model, computed across prompt variants. Higher values indicate greater intra-class predictive uniformity.	40
3.6	Examples where none of the models predicted the gold label.	46
3.7	Representative examples from the 198 high-sensitivity cases. Relabeled cases reflect corrected annotation inconsistencies; Removed cases lacked sufficient semantic content; Kept cases illustrate borderline inference distinctions.	48
3.8	Pairwise Cohen’s Kappa scores across all prompt variants, indicating the degree of agreement between model predictions. Higher values reflect stronger alignment in label choices across the full dataset. Claude and GPT-4o exhibit the highest agreement, whereas Aya-23-35B shows lower agreement with the other models.	49
3.9	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.	50
4.1	Label distribution across synthetic and gold datasets.	61
4.2	Lexical statistics across synthetic dataset sizes and gold data. TTR decreases with size as a natural corpus-size effect. Unique tokens grow steadily, confirming vocabulary expansion at each scale. Gold coverage reports the proportion of gold vocabulary present in the synthetic data. .	62
4.3	Distribution of labels across Fluent and Non-fluent data	68
4.4	Parameter counts for mT5-small and mT5-base in both encoder-decoder and encoder-only variant with classification head	70
4.5	Performance of mT5-small across different dataset sizes (Gold, Synthetic only, Synthetic+Gold, Synthetic→Gold, Gold →Synthetic). Scores are reported as mean ± 95% CI. Best results per section are highlighted in bold.	75
4.6	Performance of mT5-base across different dataset sizes (Gold, Synthetic only, Synthetic+Gold, Synthetic→Gold, Gold →Synthetic). Scores are reported as mean ± 95% CI. Best results per section are highlighted in bold.	76
4.7	Comparison with ParsiNLU benchmark. Δ indicates improvement over the gold-only baseline for each model.	77
5.1	Label distribution across synthetic datasets generated by different annotator configurations. Notable variations in class balance reflect model-specific annotation biases.	87

5.2	Accuracy of different models under prompt variations with Prompts-ensemble on the ParsiNLU (train) dataset. Results are reported for reverse/original, elaborate/concise, and English/Persian prompts. The last column shows ensemble accuracy using Models-ensemble with a random tie-break. The number of tied cases is 63, 60, 50, 51, 43, and 32, respectively. Best individual prompt scores are underlined, and the overall best scores are highlighted.	90
5.3	mT5-small performance across annotator quality spectrum and training regimes using 50k synthetic data. Results demonstrate the impact of annotation quality from random labels to ensemble methods across three training scenarios: synthetic-only, combined synthetic+gold, and synthetic pre-training followed by gold fine-tuning. Performance metrics (ROC-AUC, Accuracy, Weighted F1) are reported as mean $\pm$ 95% confidence intervals, revealing systematic quality gradients from weakest to strongest annotators.	95
5.4	mT5-base performance across annotator quality spectrum and training regimes using 50k synthetic data. Results demonstrate the impact of annotation quality, from random labels to ensemble methods, across three training scenarios: synthetic-only, synthetic+gold, and synthetic pre-training followed by gold fine-tuning. Performance metrics (ROC-AUC, Accuracy, Weighted F1) are shown as mean $\pm$ 95% confidence intervals, revealing systematic quality gradients from weakest to strongest annotators.	96
5.5	Class-wise mT5-small performance (mean $\pm$ 95% CI) after synthetic(Ours) $\rightarrow$ gold training, showing annotator quality effects across NLI classes. Aya-worst vs. Aya-best, GPT-4o-mini, a DeBERTa baseline, and ensembles. Higher-quality annotators (Aya-best, GPT-4o-mini) yield consistent gains—most pronounced for <i>Entailment</i> —while <i>Neutral</i> remains comparatively difficult across models.	100
5.6	Class-wise mT5-base performance (mean $\pm$ 95% CI) after synthetic(Ours) $\rightarrow$ gold training, showing annotator quality effects across NLI classes. Aya-worst vs. Aya-best, GPT-4o-mini, a DeBERTa baseline, and ensembles. Higher-quality annotators (Aya-best, GPT-4o-mini) yield consistent gains—most pronounced for <i>Entailment</i> —while <i>Neutral</i> remains comparatively difficult across models.	100

1.1 Motivation

The recent rise in the capabilities of Large Language Models (LLMs) has led to their widespread use in high-resource languages. This growth has motivated efforts to extend their use to low-resource languages. These models' success depends on access to massive, high-quality datasets with billions of tokens [63, 53] and substantial computational resources. The need for large-scale data has created a barrier: the benefits of LLMs remain concentrated in high-resource languages, especially English. Most existing foundation models are trained predominantly on English and a few major languages, leaving hundreds of languages underrepresented or entirely excluded [61]. This linguistic imbalance creates a digital divide in AI accessibility and performance, limiting the usefulness of these technologies for low-resource language speakers [94, 98, 116]. While LLMs can generalize to some extent via few-shot or in-context learning, their accuracy and reliability remain limited without fine-tuning on task-specific gold-standard datasets. This limitation is particularly evident in low-resource languages, where even advanced models such as GPT-4 exhibit substantially stronger performance in English than in these languages [95, 71]. While recent proprietary models from providers such as OpenAI and Anthropic have improved multilingual performance, their high cost and restricted access pose challenges for widespread adoption and reproducibility. In contrast, open-weight models offer greater accessibility but often lag in performance for low-resource languages. Additionally, the high computational cost, latency, and resource requirements of deploying general-purpose LLMs motivate the development of smaller specialized models that are fine-tuned on curated labeled datasets. Hence, data for fine-tuning and adaptation remains crucial for the use of LLMs beyond dominant languages.

Although researchers have developed human-curated datasets for fine-tuning in many languages [127], creating new datasets remains particularly expensive and time-consuming for languages with limited annotator availability. These constraints prevent effective

scaling of dataset size, creating a bottleneck for multilingual model development. To address this limitation, translation has been widely adopted. Many multilingual datasets, such as XNLI, XQuAD, and MASSIVE [25, 10, 32], are derived from translations of established English datasets. While these resources provide valuable benchmarks, they often introduce artifacts that do not reflect authentic language use in the target languages. This can lead to inflated model performance estimates and limited generalizability in real-world, native contexts [9, 18].

Recent advances in LLMs open a new avenue for scalable and human-like data creation. Beyond their capacity to solve tasks, LLMs have increasingly been employed as tools for creating and labeling data [132, 77, 82]. However, ensuring the quality of synthetic data remains a critical challenge. LLMs may produce hallucinations, inconsistencies, or factual errors [80, 75, 131], which increases the need for human evaluation and thus hinders scalability. To address this limitation, researchers have explored the use of LLMs as automatic evaluators, or LLM-as-a-judge [157, 78, 74]. In this framework, one or more LLMs are tasked with assessing the quality, correctness, or preference ranking of generated outputs, enabling faster and more reproducible evaluation pipelines. However, the reliability of LLMs as alternatives to human annotators remains an open question due to issues such as sensitivity to prompt variations, biases inherited from training data, weaker performance on low-resource languages, difficulties with subjective or nuanced tasks, inconsistent annotation formats, and the risk of propagating errors that may affect downstream analyses [100].

Given these limitations, data generated and annotated by LLMs can contain noise, inconsistencies, and systematic errors. While prior research in machine learning [14] shows that weak or noisy supervision can still improve performance in low-data regimes, it remains unclear whether this principle holds when the supervising model itself performs poorly in the target language. This thesis examines whether the generation and annotation capabilities of LLMs—despite their limitations—can provide sufficiently informative training signals in data-scarce settings and under what conditions synthetic data contributes positively to model performance. To this end, the thesis develops a scalable, reliable data-generation pipeline that integrates LLM-based generation and annotation. To evaluate the effectiveness of the pipeline, this thesis specifically focuses on the Natural Language Inference (NLI) task—the task of determining whether a hypothesis is entailed by, contradicted by, or neither supported nor contradicted by a premise—as a foundational task of semantic understanding [23, 26]. Data scarcity is particularly acute for NLI, where data creation and annotation require fine-grained linguistic judgment [46, 99]. Farsi, spoken by over 110 million people, highlights the scale of this gap: the largest available Farsi NLI dataset contains only around 10,000 examples [8, 66], compared to 570,000 examples for English SNLI [15]. The resulting digital divide limits the applicability of

NLP advances to speakers of low-resource languages.

While there is a growing body of research on AI-powered synthetic data generation to address the data scarcity bottleneck, the area is underexplored for low-resource languages such as **Farsi (Persian)**. Farsi presents a unique combination of linguistic and orthographic characteristics that distinguish it from many well-studied languages. While it uses a script derived from Arabic, Farsi is, in fact, an Indo-European language [85]. By focusing on Farsi, this research contributes to linguistic inclusivity and advances the understanding of how synthetic and LLM-labeled data can improve NLP systems in underrepresented languages. The next section outlines the specific research problem and objectives that this thesis seeks to address.

1.2 Research Questions

The primary goal of this thesis is to examine the role of LLMs in synthetic data generation and annotation, specifically within low-resource language scenarios. To achieve this, the following research questions are posed:

RQ1: How reliable are LLMs when used as automatic annotators in Farsi NLI?

This research question evaluates the reliability and robustness of LLMs when employed as automatic annotators in low-resource language contexts. Although LLMs can generate labeled data with minimal human input, their annotations often vary depending on prompt formulation, language choice, and model family. Understanding the extent of this sensitivity and identifying stable, reproducible annotation settings is critical for using LLMs as dependable labeling tools.

RQ2: Can an LLM-based process generate synthetic data that improves the performance of NLP models on the NLI task in Farsi? This research question assesses the effectiveness of synthetic data generated and labeled by LLMs as a complement to, or potential alternative for, limited human-annotated datasets. Current multilingual models show limited generalization when trained on small human-curated corpora. A systematic evaluation of a synthetic data pipeline that integrates generation and labeling is necessary to determine whether automatically generated data can meaningfully enhance NLI performance in Farsi.

RQ3: To what extent does the annotation quality of the LLM labeler affect downstream NLI task performance? This research question examines the relationship between the labeling quality of LLM annotators and the utility of the resulting syn-

thetic data for downstream model training. Prior work on self-training and co-training has demonstrated that noisy labels can improve downstream performance under certain conditions, such as when label noise is random rather than systematic, when sufficient data volume is available, or when it is combined with small amounts of clean data [14]. However, the quantitative relationship between LLM labeling reliability and downstream gains remains underexplored. By systematically varying prompt configurations across best, worst, random, and ensemble settings, this thesis quantifies how annotation reliability translates into downstream performance gains, revealing the practical trade-offs between labeling cost and model utility.

1.3 Thesis Overview

This thesis comprises six chapters structured around two interrelated threads: (i) the evaluation of LLMs as annotators and (ii) the generation of synthetic training data using LLMs. Below is a brief overview of each chapter:

Chapter 2 reviews the prior research on large language models, emphasizing the persistent limitations faced by low-resource languages. It situates recent developments in LLM-driven synthetic data generation and annotation within this landscape. Finally, it positions NLI as a downstream-task case study.

Chapter 3 addresses RQ1 by evaluating the reliability of LLMs as annotators of the NLI task in Farsi. Using a diverse set of open and proprietary models across multiple prompt variations, we analyze the zero-shot performance of LLMs along several dimensions: consistency, sensitivity, overall annotation accuracy, and per-class classification patterns in comparison to human labels.

Chapter 4 addresses RQ2 and introduces a modular pipeline that generates synthetic data for the NLI task in Farsi. The pipeline consists of a generative stage and an evaluative stage, where a model assigns NLI labels. The generated synthetic data is used together with the gold-standard dataset to train models under three different training regimes: (1) fine-tuning solely on the synthetic dataset; (2) fine-tuning on a merged set of synthetic and gold-standard data; and (3) warming up the model on synthetic data followed by further fine-tuning on the gold data.

Chapter 5 addresses RQ3 by exploring whether differences in annotator quality translate into differences in downstream performance for student models trained on synthetic data. To this end, datasets labeled by LLMs under best, worst, and random label configu-

rations are systematically compared, and ensemble labeling strategies combining multiple prompts or models are additionally explored. Each labeled dataset is used to fine-tune the same NLI model, enabling a controlled comparison of downstream performance relative to labeling quality. The chapter further analyzes the correlation between prompt accuracy and label agreement with ensemble outputs, assessing how annotation reliability translates into downstream-task performance.

Chapter 6 concludes the thesis by discussing its contributions, limitations, and potential future work for LLM-based data generation in low-resource language scenarios.

1.4 Contributions

This thesis makes the following contributions:

- It introduces a modular pipeline for large-scale synthetic NLI data generation and annotation in Farsi, enabling scalable training data generation for low-resource NLP tasks.
- It presents an in-depth behavioral study of LLMs as annotators for NLI, systematically evaluating the impact of prompt phrasing, language selection, and model family.
- It provides a comprehensive evaluation of robustness and agreement metrics and explores ensemble strategies to enhance annotation reliability in Farsi NLI.
- It quantifies the relationship between annotation quality and downstream model performance.

Parts of the work presented in this thesis have been published in the following peer-reviewed venue:

- **Panahi, S.**, Kelleher, J., & Nedumpozhimana, V. *SynthLLM: An LLM-based Scalable Synthetic Data Generation Pipeline for Natural Language Inference in Farsi*. LREC, 2026 (Corresponds to Chapter 4.)

Overall, this thesis demonstrates that LLMs offer a promising approach to addressing data scarcity in low-resource languages such as Farsi. The results show that synthetic data generated and labeled by LLMs can considerably enhance the performance of downstream NLP models, particularly when used in a sequential training regime (synthetic pre-training followed by gold fine-tuning). However, this potential is tempered by several key sensitivities: LLM performance is highly dependent on prompt design, model family, and the inherent subjectivity of tasks such as NLI. While the findings confirm established

1.4. CONTRIBUTIONS

principles of semi-supervised learning—namely, that label quality is critical and that ensemble methods improve reliability—this thesis’s contribution lies in quantifying these relationships specifically for LLM-based synthetic data generation.

2

Background and Related Work

This chapter introduces the background and related work, highlighting the open problems that motivate the thesis.

2.1 The Linguistic Divide in NLP: From Multilingual Models to General-Purpose LLMs

The concentration of NLP resources in high-resource languages has created a digital divide in language technology. Multilingual language models were developed to extend language technology to a broader range of the world’s languages. Models such as mBERT, XLM-R, mBART, mT5, BLOOM, and XGLM [152, 24, 81, 155, 151, 79] have been pre-trained on over 100 languages simultaneously and are widely used for NLP tasks in low-resource languages [107]. However, spreading fixed model capacity across many languages creates a trade-off known as the curse of multilinguality: adding more languages tends to degrade per-language representation quality, particularly for low-resource ones [24]. Moreover, the training data of these models remains heavily skewed toward a few dominant languages. As a result, performance on low-resource languages often lags behind that of high-resource ones, frequently necessitating task-specific fine-tuning to close this gap [73, 7]. The emergence of general-purpose, decoder-only large language models with billions of parameters, such as GPT-3 [16], has led the NLP community to shift its focus from task-specific fine-tuning toward prompt engineering. In prompt-based approaches, pre-trained LLMs are prompted using either task instructions alone—known as zero-shot prompting—or task instructions paired with demonstrations—known as few-shot prompting—to perform downstream tasks without any parameter updates. During post-training, these models are trained via supervised fine-tuning (SFT) on instruction-response pairs. After instruction tuning, LLMs have demonstrated a remarkable ability to generalize to unseen tasks [117, 146, 21]. Extending this paradigm to multilingual settings, models such as mT0, BLOOMz, and Aya represent explicitly multilingual instruction-tuned variants,

fine-tuned on multilingual task mixtures designed to support zero-shot transfer across a broad range of languages [91, 137]. Beyond instruction tuning, models are further aligned with human preferences and intentions through preference-optimization methods. This alignment stage has evolved from reinforcement learning from human feedback (RLHF) toward more efficient methods such as Direct Preference Optimization (DPO), and more recently, toward reinforcement learning with verifiable rewards such as Group Relative Policy Optimization (GRPO) [70, 108, 126]. These techniques underpin many of today’s most capable models.

Despite advances in multilingual pre-training, model scaling, and instruction tuning, both multilingual pre-trained models and general-purpose LLMs remain inherently dependent on the availability of training data. Multilinguality, scaling, and prompt-based generalization do not resolve the fundamental bottleneck: the scarcity of labeled data in low-resource languages.

2.2 From Human to LLM-based Annotation

Supervised NLP systems depend on annotated data, yet producing high-quality annotations at scale remains one of the most persistent bottlenecks in the field. Manual annotation—whether carried out by domain experts or crowd workers—is expensive, time-consuming, and challenging to scale [105, 128]. Even with careful guidelines, inter-annotator agreement on tasks requiring semantic judgment is often moderate at best, because annotators bring different linguistic intuitions and background knowledge to it [11, 118, 38]. These difficulties are compounded for low-resource languages, where qualified annotators are scarce, crowdsourcing platforms offer limited language coverage, and the cultural assumptions embedded in annotation guidelines—typically designed for English—may not transfer to other languages [61]. As a result, the NLP community has pursued a succession of strategies to reduce dependence on manual annotation, progressing broadly from translation-based projection and surface-level data augmentation, through weak supervision, to the most recent paradigm of LLM-based annotation generation.

Translation-based projection. One of the earliest and most widespread responses to data scarcity in non-English languages has been to translate existing English datasets, either automatically or via professional translators. This strategy underlies many widely used multilingual benchmarks: XNLI, XQuAD, and XCOPA [25, 10, 103] were all constructed by translating English source data. While this approach is efficient and enables rapid cross-lingual coverage, it introduces systematic biases. Translated text retains the syntactic patterns and lexical choices of the source language—a phenomenon known as *translationese*. This artificially inflates evaluation scores for models trained on similar

2.2. FROM HUMAN TO LLM-BASED ANNOTATION

translated data without reflecting genuine target-language competence [9]. Moreover, cultural assumptions embedded in the original English data —such as Western-centric commonsense or crowd-annotator biases—propagate into the translated datasets unchanged [118]. Most importantly for classification tasks, label fidelity can degrade during translation: entailment relations that depend on pragmatic implicature, cultural knowledge, or language-specific constructions may not survive the translation process intact [103]. For NLI specifically, this is an important concern: a premise–hypothesis pair that encodes Entailment in English may shift to Neutral or Contradiction once translated if the inference relies on language-specific lexical semantics or cultural world knowledge that does not carry over to the target language.

Data augmentation and weak supervision. A range of techniques has been proposed to maximize the utility of limited labeled data. Surface-level data augmentation applies transformations such as synonym replacement, random insertion and deletion, and back-translation paraphrasing to existing examples, generating new training instances without additional human annotation [148, 122]. While effective for increasing volume, these methods produce examples that remain distributionally close to the originals, offering limited lexical and syntactic diversity and rarely introducing genuinely novel linguistic patterns. A related line of work uses heuristic or programmatic labeling—often called *weak supervision*—to generate large quantities of noisy labeled data from rules, knowledge bases, or pattern-matching functions [112]. These approaches can scale efficiently, but the resulting labels tend to be noisy and domain-specific, and their effectiveness depends heavily on the availability of appropriate knowledge sources—which are often limited for low-resource languages. Self-training and co-training extend this idea by leveraging automatically generated pseudo-labels to expand the training set. *Self-training* [153] iteratively augments the labeled data using a model’s own high-confidence predictions, while *co-training* [14] trains two models on complementary feature views, each providing pseudo-labels for unlabeled data used by the other. These methods demonstrate that noisy or automatically generated labels can improve downstream performance, provided the label noise is sufficiently random rather than systematic; systematic noise, by contrast, reinforces consistent errors.

Heuristic dataset construction. Beyond translation and augmentation, several heuristic and semi-automatic methods have been used to create training resources [48, 43, 67]. For example, summarization corpora such as XLSum [48] were built by automatically pairing news articles with their summaries. However, quality varies sharply across languages: in Farsi, only 49% of XLSum summaries were judged factually consistent with their source texts [4], highlighting the limitations of automated pipelines applied uniformly

2.3. CHALLENGES AND LIMITATIONS OF LLM-BASED DATA SYNTHESIS

across diverse languages.

LLM-based data generation and annotation. The emergence of instruction-tuned LLMs has enabled a qualitatively different approach to dataset creation. Unlike the methods described above, which transform, perturb, or heuristically label existing text, LLMs can generate new text conditioned on task-specific instructions and assign labels to it in the same pipeline. Two distinct uses of LLMs have emerged within this paradigm, and it is important to separate them because they introduce different failure modes. The first is *LLM as generator*: using a model to produce new training examples from scratch. Wang et al. [144] introduced SELF-INSTRUCT, a bootstrapping procedure in which a language model generates its own instruction-response pairs, which are then filtered and used for fine-tuning. Taori et al. [132] applied this approach to distill GPT-3.5 outputs into a smaller Llama model, demonstrating that a modest set of LLM-generated instructions could produce a competitive instruction-following model at a fraction of the cost. Subsequent work extended generation to domain-specific settings: Gunasekar et al. [45] showed that curated synthetic textbook-style data could train a small model to achieve strong coding performance, while Honovich et al. [54] generated diverse task instructions at scale to broaden coverage beyond manually curated datasets. The second is *LLM as annotator*: using a model to assign labels to existing or generated data. Gilardi et al. [40] compared GPT-4 against trained crowd workers across several classification tasks and found that GPT-4 matched or exceeded crowd-worker agreement on most tasks, at substantially lower cost. He et al. [51] formalized this into ANNOLLM, a framework for using LLMs as annotation assistants with structured prompting.

In practice, the two roles—generation and annotation—are often combined. This combination can be understood as a form of knowledge distillation, in which a capable teacher model produces supervision signals that are used to train a smaller student model [154, 62]. Gekhman et al. [37] demonstrated this approach concretely. They used an LLM to annotate model-generated summaries for factual consistency, producing synthetic training data (TrueTeacher) that outperformed rule-based perturbation of human-written text. Several recent surveys provide broader coverage of this rapidly growing area [138, 131].

2.3 Challenges and Limitations of LLM-based Data Synthesis

While LLM-based data synthesis offers clear advantages in cost and scalability, the approach introduces reliability concerns at both stages of the pipeline: data generation and annotation. Crucially, these concerns are not independent: noise introduced during generation is inherited and potentially amplified during annotation, and both are exacerbated

2.3. CHALLENGES AND LIMITATIONS OF LLM-BASED DATA SYNTHESIS

when the target language is underrepresented in the model’s pre-training data [131, 80]. These concerns can be broadly categorized into systematic biases, methodological instabilities, task-specific constraints, and low-resource language challenges:

Generation quality and distributional bias. When LLMs are used to generate training text, the resulting data reflects the model’s pre-training distribution rather than the true distribution of the target language or domain. Two failure modes are particularly consequential for downstream training quality. First, *hallucination*: LLMs can produce text that is fluent and coherent on the surface but factually incorrect or semantically flawed. This problem is well-documented for summarization and question answering [59, 86] and extends to any generative data pipeline. Second, *distributional narrowness*: LLM-generated text can exhibit limited diversity, with repeated or homogeneous examples arising from fixed prompts and a small pool of in-context demonstrations [76]. Such homogeneity undermines the effectiveness of synthetic training data, especially for tasks such as sentiment analysis or natural language inference, where robustness to diverse registers and writing styles is essential.

Methodological Instability. LLM responses are highly sensitive to minor changes in prompt phrasing, formatting, and exemplar selection [121, 82]. As a result, a prompt that yields high agreement with human judges in one formulation can produce substantially degraded performance when paraphrased, even if the intended meaning is preserved. Beyond format, the *ordering* of information within a prompt has also been shown to alter model reasoning: Chen et al. [19] demonstrated that reversing the order of premises in reasoning tasks changed model predictions substantially, and Guan et al. [44] documented similar order effects across diverse reasoning tasks. This sensitivity extends to decoding parameters such as temperature and top-p sampling, which can introduce unwanted randomness into annotation tasks where consistency and reproducibility are expected. This instability is further compounded in multilingual settings, where prompt language and translation strategy have been shown to affect performance substantially: Mondshine et al. [90] found that prompting in the target language versus English yields different results, while Gan et al. [36] showed that language-specific prompt templates produce inconsistent behavior across typologically diverse languages, suggesting the effect is not limited to language choice alone. This instability necessitates a systematic investigation into which prompting and configuration strategies can yield stable, reproducible annotations for low-resource languages.

Task complexity and systematic biases. Beyond prompt sensitivity, task complexity and systematic biases introduce further reliability concerns. On relatively straight-

forward classification tasks—such as topic labeling and sentiment detection—LLMs can achieve annotation quality comparable to, or in some cases exceeding, that of crowd workers [40, 47]. However, performance degrades markedly on tasks that require pragmatic reasoning, world knowledge, or the resolution of genuine ambiguity [89, 17, 141]. Beyond task difficulty, LLMs are prone to adopting shallow heuristics. For example, models may rely on lexical overlap, tending to predict entailment when the premise and hypothesis share substantial surface-level vocabulary [87]. They may also exhibit veridicality biases, whereby models tend to assume that certain linguistic contexts are veridical even when such inferences are not warranted by the text [114]. Models can also exhibit positional bias, prioritizing items based on their position within a given prompt rather than their content [119]. For annotation tasks, these biases are especially concerning because they can generate *systematic* rather than random errors. Such annotation noise can negatively affect the generalization performance of downstream models [129].

The Low-Resource Disadvantage. The limitations described above are not merely replicated but *amplified* in low-resource settings. LLMs are generally weaker in low-resource languages due to an imbalance in pre-training data [35, 58, 149]. This creates a double bind: the communities most in need of scalable annotation solutions are those for whom the current technology is least reliable.

Surveys and empirical studies consistently synthesize these axes, concluding that the reliability of LLMs as annotators is profoundly task- and setup-dependent, calling for standardized protocols and benchmarks for using LLM-as-annotator, along with careful implementation and awareness of their limitations [136, 20, 74, 78, 131]. The relevant research question is no longer whether LLMs can replace human annotators, but where and how they can be used reliably.

2.4 Farsi NLP

2.4.1 Linguistic Challenges

Farsi (also referred to as Persian) is spoken by more than 110 million people across Iran, Afghanistan (as Dari), and Tajikistan (as Tajik). From a typological standpoint, Farsi sits at an interesting intersection. Linguistically, it belongs to the Indo-European family—a branch that is grammatically distant from Semitic languages—yet it is written in a modified Arabic script and absorbs a large proportion of its vocabulary from Arabic [85]. Farsi presents a distinct combination of properties that compound the difficulties faced by NLP systems designed primarily for English and other high-resource languages. First, Farsi is written right-to-left in a modified Perso-Arabic script with no capitalization, positional glyph variants, and unwritten short vowels—meaning a single written form can

correspond to multiple spoken words with distinct meanings. The zero-width non-joiner (ZWNJ) character, used to separate morphemes without a visible space, is inconsistently applied in real corpora, routinely corrupting tokenization pipelines [106]. These script-level properties interact poorly with subword tokenizers such as BPE [123] and SentencePiece [69], which are trained predominantly on Latin-script data and over-segment Farsi words relative to English. As shown in Figure 2.1, a semantically equivalent Farsi sentence is tokenized into 29 subword tokens compared to 16 for its English counterpart^a. Crucially, this over-segmentation is not merely a matter of computational cost or sequence length—it carries a semantic cost. When a morphologically coherent unit in Farsi, such as an inflected verb or a suffixed noun, is split into several subword fragments, each fragment, in isolation, encodes only partial morphological information, and the model must reconstruct the intended meaning from pieces that may be severely underrepresented in the training data. In low-resource settings, where such combinations are already rare, this fragmentation amplifies token-level data sparsity, ultimately degrading the model’s ability to form coherent semantic representations [115, 5].



Figure 2.1: Tokenization comparison of semantically equivalent English and Farsi text. The English sentence consists of 16 tokens, while the Farsi (right-to-left) sentence yields 29 tokens, highlighting differences in script and tokenization behavior across languages.

At the morphological level, Farsi is highly inflected: verbs inflect for person, number, tense, aspect, and mood; nouns carry possessive, plural, and object-marking suffixes.

^a<https://platform.openai.com/tokenizer>

A single lemma may surface in dozens of morphologically distinct forms, resulting in severe vocabulary sparsity on small corpora and limiting cross-inflection generalization [1]. Structurally, Farsi follows a subject–object–verb base order, though constituent ordering is relatively free due to its reliance on morphological case and agreement cues rather than positional syntax. Cross-lingual transfer has been shown to degrade when source and target languages differ in basic word order, as positional heuristics implicitly learned from English pre-training data—such as expecting the verb between subject and object—do not transfer to SOV languages [6]. For NLI in particular, where Entailment and Contradiction often hinge on exactly who is doing what to whom, such structural mismatches can cause models to misattribute thematic roles, systematically misleading Entailment and Contradiction judgments.

At the lexical level, Farsi draws heavily on Arabic in formal registers [101]. This lexical heterogeneity makes it challenging for NLP systems to perform tasks such as tokenization, part-of-speech tagging, and word-sense disambiguation, as models must distinguish between native Persian, Arabic-derived stems, and foreign borrowings [125].

2.4.2 Farsi as a Representative of Low-Resource Languages

The term *low-resource* in NLP typically denotes a language for which publicly available annotated data is substantially scarcer than for English, often by one or more orders of magnitude depending on the task. Farsi has over 1 million Wikipedia articles^b and has been classified as a language with a strong web presence but insufficient efforts in labeled data collection, placing it in an intermediate tier between high-resource and severely low-resource languages. Despite these resources, research on subjective tasks in Farsi remains notably scarce, making it a low-resource language in this domain [113]. Table 2.1 summarizes the existing Farsi datasets, including both natively developed corpora and those produced via translation.

To illustrate the disparity, the largest Farsi NLI dataset (FarsTail) contains fewer than 10,000 examples [8], compared with 570,000 for English SNLI [15]. FarsiQuAD is a notable exception: with approximately 145,000 question–answer pairs [34], approaching the scale of English SQuAD 2.0’s 150,000 questions [110]. Machine translation has benefited from sustained parallel corpus curation, though quality remains uneven across domains and registers. Pre-training data for multilingual models reflects the same disparity: for example, the mC4 dataset contains approximately 365 million rows for English but only 54 million for Farsi, meaning that multilingual models receive substantially less exposure to Farsi linguistic patterns during pre-training, with direct consequences for downstream task performance [155].

^bhttps://en.wikipedia.org/wiki/Persian_Wikipedia

Dataset	Task	Size	Type	Year	Ref
Monolingual Farsi — Natural Language Inference					
ParsiNLU (NLI)	NLI	~2.7k	Native / Trans.	2021	[66]
FarsTail	NLI	10k	Native (semi)	2023	[8]
Monolingual Farsi — Question Answering					
ParsiNLU (QA)	QA / RC	~14.5k	Native	2021	[66]
PQuAD	Extractive QA	~80k	Native (Wiki)	2023	[27]
PersianQA	Extractive QA	~9k	Translated	2021	[13]
FarsiQuAD	Extractive QA	~145k	Native (Wiki)	2024	[34]
Monolingual Farsi — Sentiment & Emotion					
SentiPers	Sentiment	~15.5k	Native (reviews)	2018	[56]
SnappFood	Sentiment	~70k	Native (reviews)	2020	[134]
Monolingual Farsi — Named Entity Recognition					
ARMAN	NER	250k tok / 7.7k sent	Native (news)	2018	[104]
Peyma	NER	~302k tok	Native (web)	2018	[124]
ParsTwiNER	NER	~17k tweets	Native (Twitter)	2021	[3]
Monolingual Farsi—Summarization & MT					
XL-Sum	Summarization	~94k	Native (BBC)	2021	[48]
MIZAN	MT (FA–EN)	~1M	Parallel	2016	[64]
Multilingual (Farsi subset)					
mC4	LM pre-train	~52B tok	Web crawl	2021	[155]
MASSIVE	NLU	~19.5k	Translated	2022	[32]

Table 2.1: Overview of Farsi NLP datasets. "Type" indicates whether data is natively authored or derived via translation/crawling.

While recent benchmarking initiatives have begun to establish evaluation standards for Farsi—ParsiNLU [66] for general NLU, ParsBench [97] for LLM evaluation, and the Khayyam Challenge (PersianMMLU) [39] for culturally grounded QA—Farsi lacks the recurring shared tasks and community-driven annotation efforts that have driven rapid progress in English NLP. This combination of limited labeled data, underdeveloped evaluation infrastructure, and typological distance from English makes Farsi a particularly informative case study for investigating synthetic data generation methods.

2.4.3 State of the Art Across Farsi NLP Tasks

As of the time of writing, the strongest results on Farsi NLP tasks are consistently achieved by fine-tuning multilingual pre-trained models on the available task-specific data. A consistent pattern emerges across tasks: On tasks with larger, more objectively annotated datasets, performance is strong: sentiment classification on SnappFood and SentiPers reaches F1 scores above 85% with ParsBERT [30], NER on the ARMAN and Peyma

corpora achieves comparable results with XLM-RoBERTa [130], and PQuAD supports competitive extractive question answering with fine-tuned transformer models (74.8% EM / 87.6% F1) [27]. However, for tasks requiring reasoning over sentence pairs, performance lags substantially. The best reported accuracies on FarsTail and ParsiNLU are significantly lower than those in English, reflecting both smaller training sets and the compounding effects of the linguistic challenges described above. Recent work on Farsi stance detection provides further evidence of this challenge. Zarharan et al. [156] introduce FarExStance, a large-scale explainable stance-detection dataset containing 26,307 instances with stance labels and sentence-level evidence. Despite evaluating both fine-tuned multilingual models and recent LLMs, the highest reported LLM accuracy is 79.8%, roughly matching estimated human performance of 79.2%—yet both fall short of the ceiling, indicating that a substantial portion of cases remain challenging even for humans. Recent evaluations of general-purpose LLMs confirm this gap: GPT-4 underperforms on tasks requiring nuanced semantic judgment in Farsi [2].

2.5 Natural Language Inference

Having established the resource constraints that characterize Farsi NLP, we now turn to the specific task that motivates this thesis: natural language inference. NLI, originally framed as recognizing textual entailment [26], is a pairwise classification task where the model predicts whether a premise p entails, contradicts, or is neutral toward a hypothesis h , as shown in Table 2.2. NLI is not strict logical entailment; instead, it reflects whether a human reading the premise would typically infer that the hypothesis is true.

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	Contradiction
An older and younger man smiling.	Two men smile and laugh as the cats play on the floor.	Neutral
A soccer game with multiple males playing.	Some men are playing a sport.	Entailment

Table 2.2: Examples of premise-hypothesis pairs and their corresponding NLI labels.

The ability to determine whether one sentence follows from another by inference serves as a foundational benchmark for any model claiming semantic understanding. Many NLP tasks, such as summarization, question-answering, and translation, can be reframed as entailment pairs. Smaller models fine-tuned on entailment pairs can match the few-shot performance of much larger language models on downstream classification tasks [142].

Similarly, a pre-trained NLI system can generalize across domains with only a handful of domain-specific examples, demonstrating that textual entailment functions as a unified solver for diverse NLP tasks under limited annotation budgets. Although recognizing entailment relations is not in itself sufficient for full language understanding, it remains a minimal, necessary condition for semantic competence [23, 84].

2.5.1 Benchmarks and Their Limitations

NLI has become a central benchmark for evaluating semantic reasoning, serving as a proxy for a model’s ability to move beyond surface-level lexical matching toward genuine language understanding. Large-scale datasets such as SNLI [15] and MNLI [150] have been incorporated into widely adopted evaluation suites, including GLUE [140] and SuperGLUE [139]. Yet these benchmarks suffer from several well-documented limitations: both are English-only, and human-level performance on GLUE was surpassed as early as 2019 [92, 139], raising questions about whether models were capturing genuine linguistic competence or exploiting dataset-specific artifacts [46]. In addition, both were constructed primarily from crowdsourced English data, encoding cultural and pragmatic assumptions that do not transfer cleanly to other languages.

The multilingual extension XNLI [25] attempted to address the language coverage gap by translating MultiNLI into 15 languages. However, as discussed in Section 2.2, translation-based benchmarks inherit a further layer of problems: translation artifacts can artificially inflate performance. For this reason, efforts have been made to develop native, language-specific evaluation resources that better reflect local linguistic and cultural characteristics, rather than relying on translated proxies. ParsiNLU [66] is one such initiative, providing a suite of language-understanding tasks—including NLI—tailored for Farsi and serving as an evaluation resource used in this thesis.

2.5.2 Subjectivity in NLI Annotation

Despite its structured three-way label space, NLI annotation is inherently *subjective*. Human annotators frequently disagree on the appropriate label for a given premise–hypothesis pair, particularly in cases involving pragmatic inference, world knowledge, or ambiguous phrasing [99]. These NLI cases often support multiple valid interpretations, in which identical premise–hypothesis pairs can receive different, yet justifiable, labels depending on contextual assumptions and interpretative frameworks. A key source of this subjectivity is the tension between two distinct modes of reasoning: *semantic* interpretation, which relies on literal textual meaning and logical structure, and *pragmatic* interpretation, which incorporates real-world knowledge, speaker intent, and contextual assumptions [83]. The following examples illustrate this fundamental tension between

semantic and pragmatic interpretation: ^c

Example 1

Premise: Yes, yes, it's a good thing that they do not advertise, it's really good that it's just music.

Hypothesis: It's good, it's just music, and there are no advertisements, but DJ talks a lot.

Gold: Neutral

Semantic: Extra DJ detail is not entailed but not contradictory. → Neutral.

Pragmatic: "just music" may exclude DJ talk. → Contradiction

Example 2

Premise: But some democrats say this should have been done a long time ago.

Hypothesis: Democrats say now is the time to implement this program.

Gold: Contradiction

Semantic: "Should have been done long ago" conflicts with "now is the time" because the obligation's time frame has passed. → Contradiction.

Pragmatic: the phrase can serve as a lament about delay while still supporting action now, making the statements compatible. → Entailment.

Example 3

Premise: The end user is considered a participant in the learning process and not merely a recipient of information.

Hypothesis: The end user receives the information.

Gold: Entailment

Semantic: "not merely a recipient" implies the user *is* a recipient of information, so the hypothesis follows → Entailment.

Pragmatic: The hypothesis states that the end user receives information, implying a passive role. The premise, however, holds that the end user is part of the learning process and is not merely a receiver of information, implying an active role. These two statements are in direct opposition regarding the role of the end user, leading to a Contradiction.

^cThe examples are translated from the ParsiNLU dataset [66].

These disagreements have direct consequences for data creation: NLI requires annotators to reason about implication, speaker intent, and world knowledge, making annotation substantially harder and more expensive than tasks with more objectively verifiable labels. This inherent subjectivity is reflected in consistently low inter-annotator agreement across NLI datasets [99], and it is especially consequential in low-resource settings where disagreements cannot be resolved through large annotator pools. This subjectivity presents particular challenges for LLM-based annotation, as models must navigate the same ambiguities that challenge human annotators while also contending with limited exposure to low-resource languages, sensitivity to prompt formulation, training biases, and architectural constraints that shape their judgments in ways that may not align with human reasoning. This motivates the central investigation of this thesis: a systematic examination of whether LLMs can serve as reliable annotators for Farsi NLI and under what prompting and configuration conditions their outputs best approximate human judgment.

2.5.3 Case Study: ParsiNLU

Given the considerations above, this thesis adopts the ParsiNLU dataset as its primary case study, as it represents the most carefully constructed native Farsi NLI resource currently available. The ParsiNLU benchmark [66] includes an NLI component of approximately 2,700 samples, constructed through expert annotation. This dataset was constructed via two complementary strategies:

- **Natural**—Annotators authored premise–hypothesis pairs from Farsi sources, focusing on naturally inferential sentence pairs—for example, those containing discourse markers such as *but*—alongside their preceding context sentences. Annotators were asked to consider both sentences and write a premise and corresponding entailing, contradicting, and neutral sentences. To minimize annotation artifacts, annotators were explicitly instructed to avoid trivial transformations such as negation or synonym substitution, which are known to produce spurious correlations. This approach yields high-quality instances, but it is a relatively slower annotation task.
- **MNLI**—Translating and post-editing samples from the English MNLI dataset. Furthermore, each translated document was reviewed by a native speaker annotator to correct translation errors. Notably, approximately 66.4% of translated samples required post-editing corrections, underscoring the limitations of direct translation for NLI data.

The native subset emphasizes high-quality, linguistically grounded inferences, while the translated subset provides broader topical and genre coverage. Table 2.3 shows that the dataset is relatively balanced. Figure 2.2 further illustrates how this balance shifts by

2.6. SELECTED LARGE LANGUAGE MODELS

source: MNLI-derived samples are more prevalent in the Entailment class, while naturally curated data dominates Contradiction, and the Neutral class is broadly balanced across both sources.

Split	Size	Entailment	Neutral	Contradiction
ParsiNLU (train)	750	275	241	234
ParsiNLU (validation)	269	104	87	78
ParsiNLU (test)	1668	608	559	501

Table 2.3: Label distribution of the ParsiNLU [66] dataset across train, validation, and test splits.

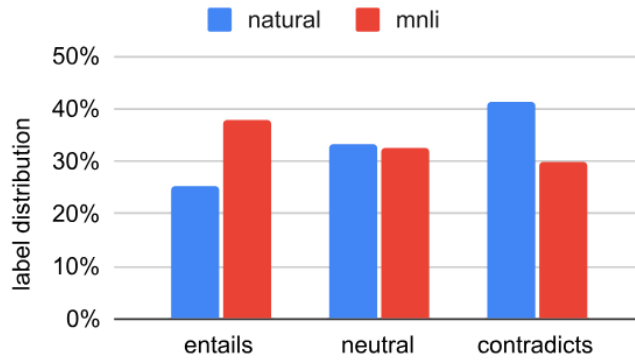


Figure 2.2: Distribution of natural and MNLI-derived samples across NLI labels in the ParsiNLU [66] dataset .

2.6 Selected Large Language Models

While LLMs, at their core, are built on transformers, they differ substantially in scale, training data, and development. To systematically investigate how underlying model families and training paradigms influence annotator reliability, we selected instruction-tuned LLMs based on access type (proprietary vs. open-weight), model scale (parameter count and computational demands), and linguistic coverage. Our selection includes three proprietary models—Claude-Opus-4^d, GPT-4o^e, GPT-4o-mini^f—which represent state-of-the-art systems with strong reasoning capabilities. In addition, we include three open-

^d<https://www.anthropic.com/news/claude-4>

^e<https://openai.com/index/hello-gpt-4o/>

^f<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>

2.7. PROMPT ENGINEERING

weight models— Aya-23-35B^g, Qwen3-30B-Instruct^h, and Llama3.3-70B-Instructⁱ—as reproducible alternatives that do not require proprietary API access^j. Table 2.4 shows the details of the models used in our experiments. To contextualize the models’ capabilities, we include its performance on the MMMLU (Multilingual Massive Multitask Language Understanding) benchmark [52]. This benchmark provides a standardized measure of knowledge across domains and languages.

Model	Size	Characteristics	MMMLU
Proprietary Models			
Claude-Opus-4	Unknown	Advanced reasoning and coding · Multimodal · Strong alignment · Multilingual	88.8
GPT-4o	Unknown	Multimodal · Strong multilingual performance	69.1
GPT-4o-mini	Unknown	Lightweight · Cost-efficient · Competitive performance	62.6
Open-weight Models			
Aya-23-35B	35B	Multilingual · 23 languages (incl. Farsi)	58.2
Qwen3-30B-Instruct	30B	Mixture-of-experts · Strong reasoning · Multilingual	78.4
Llama3.3-70B-Instruct	70B	Large-scale · General-purpose reasoning · Multilingual training	64.3

Table 2.4: Overview of LLMs evaluated in this study, including model scale, access type, and multilingual benchmark performance (MMMLU).

2.7 Prompt Engineering

A prompt is a natural-language text that asks LLMs to perform a specific task. The process of crafting an instruction to generate the desired outcome is called prompt engineering. Unlike fine-tuning, prompt engineering guides a model’s behavior without changing its weights, a process known as in-context learning [16]. Many modern LLMs follow a chat-template structure, in which interactions are formatted as a sequence of system and user messages. Models deployed via conversational APIs, such as those provided by OpenAI and Meta, process prompts using specific message roles (e.g., "system", "user", "assistant") and combine them into a structured input. "System" prompts define the

^g<https://cohere.com/research/papers/aya-command-23-8b-and-35b-technical-report-2024-05-23>

^h<https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507>

ⁱ<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

^jWe 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 in Farsi.

role of the language model, including task instructions, constraints, and response style. "User" prompts specify the task itself, providing the input content, context, or query that the model is expected to respond to. The "assistant" role represents the model's previous outputs in a multi-turn interaction. It is used to provide conversational history, allowing the model to maintain context across turns and generate coherent, context-aware responses. As our experiments involve single-turn interactions, the "assistant" role is not used. Given a system prompt and a user prompt, the model combines them into a single input sequence, which is then used to generate the output. Below is an example of a prompt with system and user roles (adopted from [57]):

Prompt

System prompt: You're an experienced real estate agent. Your job is to read the disclosure carefully, fairly assess the property's condition based on it, and help your buyer understand the risks and opportunities of the property. For each question, answer succinctly and professionally.

User prompt: Context: [disclosure.pdf]

Question: Summarize the noise complaints, if any, about this property.

Answer:

Prompting strategies vary in the degree of guidance they provide to the model. The zero-shot prompt directly instructs the model to perform a task. Few-shot prompting supplements these instructions with one or more input-output examples, providing demonstrations that have been shown to reduce ambiguity in task interpretation. The Chain-of-Thought (CoT) technique [147] encourages the model to decompose complex reasoning into a series of intermediate steps, with benefits most pronounced on complex reasoning tasks and in larger models.

While these prompting strategies can improve performance, they also highlight a key challenge: LLM outputs are highly sensitive to prompt formulation. This is a well-documented phenomenon, fundamentally linked to their autoregressive 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 [88]. Studies have documented significant order effects, where the sequence of presenting information fundamentally alters model reasoning [19, 158, 44], and have quantified their vulnerability to spurious features in prompt phrasing [121]. Beyond information ordering, the prompt template has been shown to significantly influence model performance. He et al. [50] show that different prompt templates can lead to performance differences of up to 40% for GPT-3.5-turbo on code translation tasks, while larger models such as GPT-4 are less sensitive to these variations. This instability is further compounded

in multilingual contexts, where prompt translation strategies [90] and language-specific templates [36] dramatically impact performance.

3

LLMs as Annotators for NLI

3.1 Introduction

The previous chapter established that LLM-based annotation faces compounding reliability challenges: prompt sensitivity, task-dependent performance degradation, and amplified weaknesses in low-resource languages. Before deploying LLMs as annotators in our data generation pipeline (Chapter 4), we need empirical evidence about their behavior on a semantically complex task in a genuinely low-resource setting. To this end, we evaluate six instruction-tuned LLMs as annotators for Farsi NLI across eight prompt configurations, assessing reliability through multiple complementary dimensions: overall and per-class accuracy, prompt sensitivity, intra-class consistency, and inter-model agreement. Our results show that LLM annotations for Farsi NLI are markedly sensitive to prompt design, with input-order effects producing the greatest volatility; moreover, class-wise analysis reveals that Neutral is the most challenging class for models. While proprietary models (e.g., Claude-Opus-4, GPT-4o) exhibit greater stability than open-weight models, we also find that stable prediction cases can still be confidently wrong, highlighting the need for class-aware and cross-prompt or cross-model agreement checks. The remainder of this chapter is organized as follows: In Section 3.2, we outline our experimental design and provide examples of the prompts we used. Section 3.3 presents the results and 3.4 discusses their implications. Finally, in Section 3.5, we summarize our contributions and outline directions for future research.

3.2 Method

Our methodology tests 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 the

ParsiNLU training split (750 samples: 275 Entailment, 241 Neutral, and 234 Contradiction) as our evaluation testbed. The training split was chosen instead of the test split because its size made it more manageable for running multiple experiments.

3.2.1 Prompt Design and Variations

Our prompt design framework employs controlled variations across three 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 define the role of the language model and constrain its behavior during the classification task ^a. System prompts are provided in either English or Farsi to examine the influence of instruction language on model performance.

Elaborated Instruction

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.

Concise Instruction

Classify the relationship between the following two statements as
Entailment, Contradiction, or Neutral. Respond with one word only.

We translate these two system prompts to Farsi. We do not translate the input samples (premises and hypotheses).

User Prompts To evaluate the sensitivity of language models to input phrasing, two distinct user message formulations are employed:

Input in original order

Premise:{premise} Hypothesis:{hypothesis} Answer:

This configuration presents the task in its canonical form, where the premise (providing context) precedes the hypothesis (the statement to be inferred).

^aPreliminary experiments using definitional prompts that explicitly specified the meaning of each NLI label did not yield a meaningful improvement in accuracy on a held-out subset. Consequently, these prompts were not considered further.

Input in reverse order

Hypothesis:{hypothesis} Premise:{premise} Answer:

Because Entailment is inherently directional—the premise logically precedes the hypothesis, not vice versa—we use explicit semantic role labels (Premise: and Hypothesis:) that remain attached to their respective sentences regardless of presentation order. This ensures that any performance difference between ordering conditions reflects sensitivity to surface-level sequencing rather than role confusion. We therefore treat any remaining performance differences between ordering conditions as evidence that models rely on positional heuristics that persist even in the presence of explicit semantic role labels. By combining the four system prompt variants (two in each language) with the two user prompt formats described (**original** and **reverse**), each model is evaluated across eight unique prompt configurations (Table 3.1).

Prompt	Language	Style	Order
original_elaborate_en	English	Elaborate	original
original_concise_en	English	Concise	original
original_elaborate_fa	Farsi	Elaborate	original
original_concise_fa	Farsi	Concise	original
reverse_elaborate_en	English	Elaborate	reverse
reverse_concise_en	English	Concise	reverse
reverse_elaborate_fa	Farsi	Elaborate	reverse
reverse_concise_fa	Farsi	Concise	reverse

Table 3.1: 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).

3.2.2 Post-processing and Label Normalization

Inspection of models’ outputs revealed substantial inconsistencies in model output formatting, including the use of Farsi lexical equivalents, variations in capitalization, extraneous punctuation, and labels embedded within longer explanatory outputs. To ensure evaluation robustness and comparability across models and prompts, we implemented a comprehensive label normalization procedure. This process mapped all model predictions to a canonical set of English NLI labels: **Entailment**, **Contradiction**, and **Neutral**. The normalization ensured that semantically equivalent predictions were standardized before

evaluation.

3.3 Results

To ensure a comprehensive evaluation of model behavior under prompt perturbation, we report multiple metrics that target different dimensions of robustness. Table 3.2 provides the summary of these metrics.

Metric	Granularity	Function
Accuracy	Prompt-level	Task correctness
Sensitivity	Input-level	Robustness to prompt variation
Consistency	Class-level	Stability of predictions within a class
Fleiss’s Kappa/Cohen’s Kappa	Model-level	Inter-Model agreement

Table 3.2: Multi-dimensional evaluation framework for assessing LLM annotator reliability. The taxonomy distinguishes between prompt-level, input-level, class-level, and model-level analyses to provide a holistic view of annotation stability.

3.3.1 Accuracy

We compute the model’s prediction accuracy independently for each prompt variant. This prompt-level evaluation enables a finer-grained diagnosis of prompt sensitivity. The results reveal substantial variation both across models and across prompt configurations. Accuracy ranges from 0.83 (Claude-Opus-4, `original_concise_fa`) to 0.61 (Aya-23-35B, `reverse_elaborate_fa`), a 22-percentage-point spread that shows the extent to which annotation quality depends on design choices that are often treated as arbitrary (Figure 3.1). Two patterns are immediately apparent. First, the original premise-hypothesis ordering consistently outperforms the reversed ordering for every model. This is consistent with the positional expectations that LLMs acquire during training: the vast majority of NLI training data—and indeed most premise-hypothesis formulations in the literature—present the premise before the hypothesis, so models have learned to associate this sequential structure with the task [19]. When the order is reversed, the model must reason against its learned positional prior, which degrades performance even when the semantic content is identical. Second, proprietary models exhibit a narrower, higher-accuracy band (0.74–0.83) than open-weight models (0.61–0.78), suggesting that more extensive instruction-tuning and alignment procedures confer not only higher average performance but also greater robustness to prompt variation.

One caveat applies to the absolute accuracy figures reported above. ParsiNLU has been publicly available since 2021, and we cannot rule out that portions of it appeared

3.3. RESULTS

in the pre-training corpora of the models evaluated here. The strong sensitivity of these same models to surface-level prompt perturbations—particularly to premise–hypothesis reordering and instruction language—argues against pure memorization, since memorized labels would be expected to remain stable under such perturbations. Nevertheless, the absolute accuracy figures should be read as upper bounds; the relative comparisons across models, prompts, and ordering conditions, which form the basis of the analyses that follow, are unaffected by this concern.

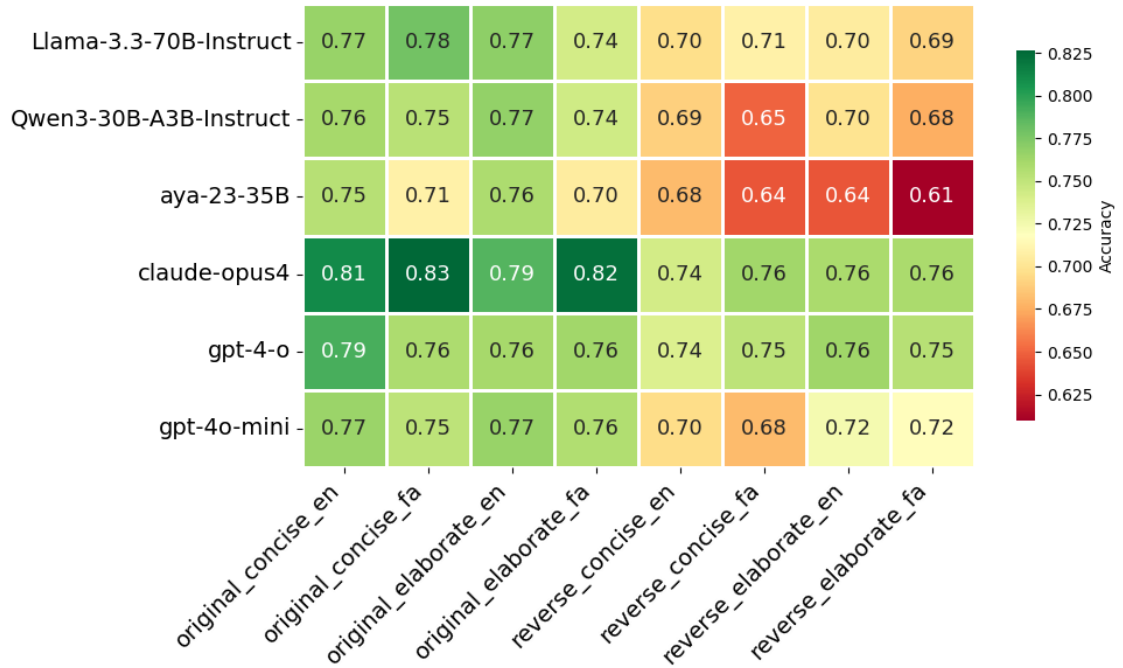


Figure 3.1: Accuracy comparison of six language models across eight prompt configurations for the NLI task. Models are evaluated under combinations of language (English/Farsi), system prompt style (detailed/concise), and premise-hypothesis ordering. The highest recorded accuracy (0.83) was achieved by Claude-Opus-4 under the `original_concise_fa` prompt, while the lowest (0.61) was observed for Aya-23-35B under the `reverse_elaborate_fa` prompt. This visualization highlights prompt and model-dependent variability in performance for the Farsi NLI annotation task.

A per-class breakdown confirms that this variability is not uniformly distributed (Figure 3.2). Entailment is classified most accurately and with the least cross-prompt variation, which is expected: Entailment pairs typically share lexical content between premise and hypothesis, providing surface-level cues that models can exploit regardless of prompt formulation. Contradiction shows moderate accuracy but greater variability, particularly among open-weight models, reflecting the additional reasoning step required to identify explicit semantic conflict. Neutral class exhibits the lowest median accuracy and the

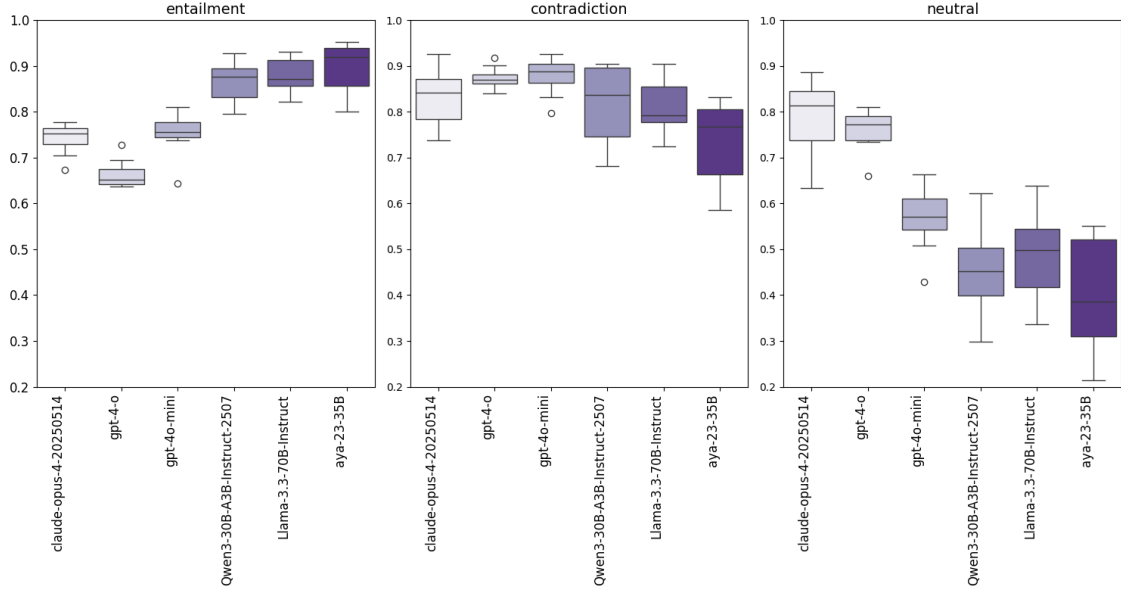


Figure 3.2: Per-Class accuracy comparison of six language models across eight prompt configurations for NLI tasks. Models are evaluated under combinations of language (English/Farsi), system prompt style (detailed/concise), and premise-hypothesis ordering. The results highlight strong and consistent performance on Entailment, greater variability and moderate accuracy for Contradiction, and substantially lower and more dispersed accuracy for Neutral, indicating that identifying Neutral relationships presents the greatest challenge for current LLMs in this setting.

largest dispersion across prompt conditions. This pattern is consistent with the inherent ambiguity of the Neutral category discussed in Section 2.5: Neutral requires the annotator to determine that a hypothesis is neither entailed nor contradicted, a judgment that depends heavily on whether the annotator adopts a strict textual reading or incorporates pragmatic inference. The difficulty LLMs face with Neutral has a direct, practical consequence for the data-generation pipeline developed in subsequent chapters. If LLM annotators systematically misclassify Neutral instances, the resulting synthetic dataset will carry a structured labeling bias that propagates to downstream model training—shown to be harmful to learned representations [129].

Evaluating a model solely on accuracy overlooks its operational reliability: a high-performing model might still be unacceptably brittle, with predictions fluctuating due to minor, inconsequential changes in the prompt. The metrics of sensitivity and consistency were introduced specifically to quantify this behavior, providing a direct measure of a model’s stability under prompt variations.

3.3.2 Prompt Sensitivity

This metric captures the extent to which a model’s predicted labels for a given input vary across semantically equivalent prompt formulations, regardless of the ground truth—such that a model that always predicts the same label, whether correct or incorrect, receives a sensitivity score of zero. Specifically, sensitivity is measured as the *normalized entropy* of the predicted label distribution across a set of prompt variants for the same input—with higher values indicating greater sensitivity [29].

Formally, for an input x and a set of prompt variants $\{p_1, \dots, p_n\}$, let $P_x \in \mathbb{R}^K$ denote the empirical distribution of predicted labels over K classes. The sensitivity for input x is then computed as:

$$\text{Sensitivity}(x) = \frac{H(P_x)}{\log_2 K}, \quad \text{where} \quad H(P_x) = - \sum_{i=1}^K P_x(i) \log_2 P_x(i) \quad (3.1)$$

For example, in the NLI task, consider a single input x evaluated under eight prompt variants, yielding the following predicted labels:

$$\{\text{E}, \text{E}, \text{E}, \text{N}, \text{N}, \text{N}, \text{N}, \text{C}\}$$

The empirical distribution over $K = 3$ classes is:

$$P_x = \left(\frac{3}{8}, \frac{4}{8}, \frac{1}{8} \right) = (0.375, 0.500, 0.125)$$

The entropy of this distribution is:

$$H(P_x) = - \left(\frac{3}{8} \log_2 \frac{3}{8} + \frac{4}{8} \log_2 \frac{4}{8} + \frac{1}{8} \log_2 \frac{1}{8} \right) = 1.406$$

Normalizing by $\log_2 K = \log_2 3 = 1.585$ gives:

$$\text{Sensitivity}(x) = \frac{1.406}{1.585} = 0.887$$

This high sensitivity score (close to 1.0) indicates that the model’s prediction for this input is highly unstable: no single class receives a clear majority of votes, and all three labels appear across different prompt formulations.

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 de-*

cision instability, particularly relevant in classification tasks such as NLI, where minor surface-level prompt variations can result in substantial shifts in model predictions.

Table 3.3 summarizes the mean sensitivity scores across all prompt variants for each model and inference class, providing a quantitative measure of how susceptible model predictions are to prompt formulation. Lower sensitivity values indicate stronger robustness and internal consistency, while higher scores reflect greater variability in the model’s predicted labels across prompts. Overall sensitivity confirms the model family divide observed in the accuracy analysis: Claude-Opus-4 (0.16) and GPT-4o (0.18) are the most robust, while Aya-23-35B (0.26) is the most volatile. However, the per-class breakdown reveals a more nuanced picture than the overall ranking suggests. The open-weight models (Aya-23-35B, Qwen3-30B, LLaMA-3.3-70B) all share a common sensitivity profile: Entailment is their most stable class (0.13–0.16), Contradiction is moderately variable (0.18–0.27), and Neutral is the most prompt-sensitive (0.33–0.36). This gradient mirrors the expected difficulty ordering: Entailment pairs offer surface-level lexical cues that anchor predictions regardless of prompt phrasing, while Neutral requires the subtler judgment discussed above. The proprietary models depart from this pattern in instructive ways. Claude-Opus-4 exhibits near-uniform sensitivity across all three classes (0.17, 0.14, 0.18), suggesting that its reasoning process is comparably stable regardless of the inferential relationship being assessed, including for Neutral, where every other model shows markedly elevated variability. This balanced profile may reflect more extensive alignment training that has regularized the model’s behavior across the full label space, rather than defaulting to high-confidence predictions only for the easier classes. GPT-4o and GPT-4o-mini display a different, unexpected pattern: both models are most stable on Contradiction (0.11 and 0.13, respectively) but least stable on Entailment (0.22 and 0.21, respectively). This is the inverse of the open-weight profile and is noteworthy because Entailment is the class that all other models find easiest to classify consistently. One possible explanation is that the GPT models’ alignment training has made them more cautious about affirming Entailment—a form of conservatism in which the model hedges on whether a hypothesis strictly follows from the premise, producing variable predictions that oscillate between Entailment and Neutral depending on prompt phrasing. By contrast, when the GPT models detect an explicit semantic conflict, they classify it as a Contradiction with high confidence and consistency. This asymmetry has practical implications for annotation pipelines: a GPT-based annotator may produce noisier Entailment labels than expected. In contrast, Contradiction labels are likely to be more reliable than those of open-weight alternatives.

Figure 3.3 illustrates the distribution of prompt sensitivity scores for all evaluated models. Each violin represents the distribution of sample-level sensitivity scores, with lower scores indicating consistent predictions across prompt variants and higher scores

3.3. RESULTS

Model	Overall	Entailment	Contradiction	Neutral
Aya-23-35B	0.26	0.16	0.27	0.36
Qwen3-30B-A3B-Instruct	0.23	0.14	0.22	0.36
LLaMA-3.3-70B-Instruct	0.21	0.13	0.18	0.33
GPT-4o-mini	0.22	0.21	0.13	0.31
GPT-4o	0.18	0.22	0.11	0.20
Claude-Opus-4	0.16	0.17	0.14	0.18

Table 3.3: Mean sensitivity scores across all prompt variants for each model, disaggregated by inference class (Entailment, Contradiction, Neutral). Lower sensitivity values indicate greater robustness to prompt reformulation, whereas higher values reflect greater variability in predictions across different prompt phrasings. Among the evaluated models, Claude-Opus-4 and GPT-4o exhibit the greatest stability, while Aya-23-35B shows the greatest prompt sensitivity—particularly for the Neutral class.

indicating 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 wider distributions, with more density at higher sensitivity scores, reflecting greater sensitivity to prompt variants. To distinguish between stable and unstable predictions, we adopt a normalized entropy threshold of 0.67. This value corresponds to the point at which a model’s vote distribution across eight prompt variants transitions from a clear class preference—where at least 75% of prompts agree on a single label—to significant inter-class confusion. Samples exceeding this threshold typically receive roughly comparable vote shares across the three NLI classes, indicating that the model lacks a stable decision boundary for that input and effectively switches between Entailment, Contradiction, and Neutral depending on how the prompt is phrased.

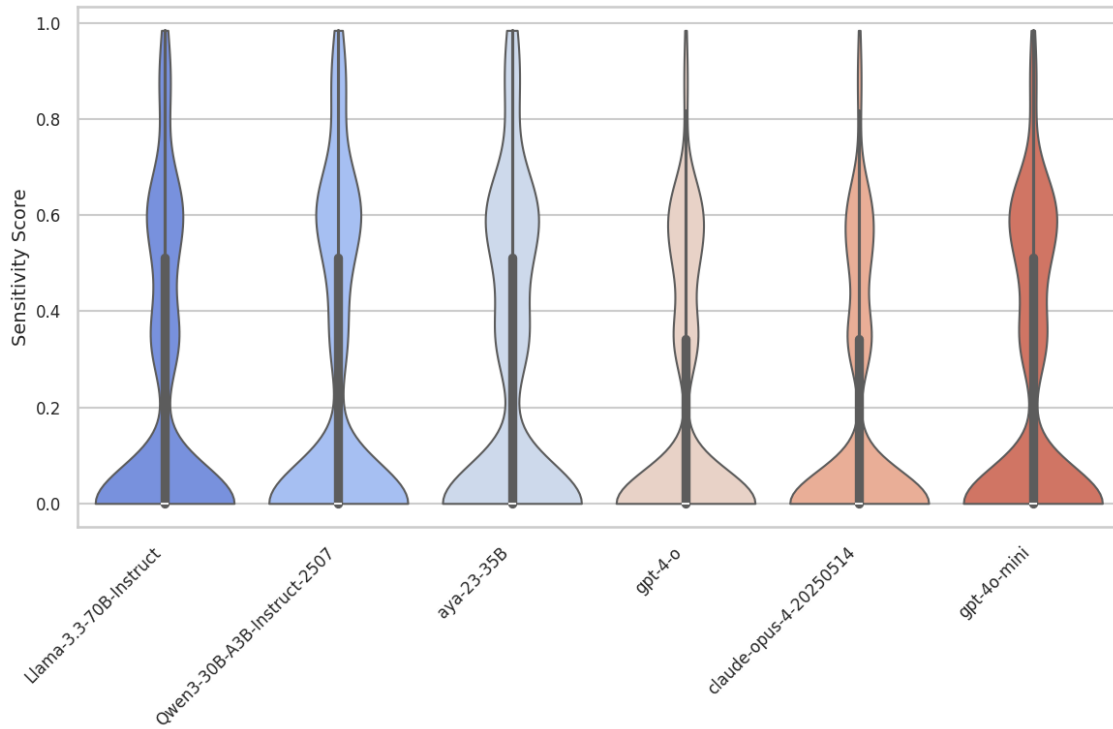


Figure 3.3: Distribution of prompt sensitivity scores across six instruction-tuned LLMs. The violin plots illustrate how each model’s predictions vary under different prompt formulations. Higher sensitivity scores indicate greater instability in responses. Proprietary models (Claude-Opus-4, GPT-4o) show 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.

To understand which factor contributes most to model sensitivity—the ordering of inputs, the language of the system prompt, or the prompt style—we recalculated sensitivities by holding one factor fixed while varying the other two. Consequently, for each factor, we evaluated sensitivity over four prompt variants. For example, under the reversed order condition, we used: `reverse_elaborate_en`, `reverse_elaborate_fa`, `reverse_concise_en`, and `reverse_concise_fa`. Table 3.4 summarizes the *mean sensitivity* of each model under three independent factors. Three patterns emerge from the factor-level comparison. The most striking is the effect of input ordering: Across all six models, the original premise-hypothesis order produces substantially lower sensitivity than the reversed order, and the gap between the two conditions is the largest of any factor pair. Claude-Opus-4 illustrates this most clearly, with sensitivity nearly doubling from 0.072 (original) to 0.123 (reversed). This confirms that positional expectations—learned from training corpora in which the premise conventionally precedes the hypothesis—represent the single strongest source of prediction instability in LLM-based NLI annotation. By

contrast, prompt language exerts a comparatively modest effect. The difference between English and Farsi system prompts is small and directionally inconsistent: Some models (GPT-4o, Aya-23-35B) are marginally more stable with English prompts, while LLaMA-3.3-70B is slightly more stable with Farsi prompts. The absence of a clear language advantage is itself notable. It suggests that, for these instruction-tuned models, the system prompt language matters less than the structural framing of the input, at least within the range of languages tested here. Prompt style (elaborate vs. concise) shows the least systematic effect. For most models, the difference between the two conditions is within 0.01–0.02, and the direction varies: LLaMA-3.3-70B is considerably more stable under concise prompts (0.157 vs. 0.206), whereas Claude-Opus-4 shows the reverse pattern (0.148 concise vs. 0.126 elaborate). This suggests that the level of instructional detail in the system prompt has a model-specific rather than a universal effect. Figures 3.4, 3.5, and 3.6 show the effect of each factor respectively.

Model name	English	Farsi	Reverse	Original	Elaborate	Concise
Aya-23-35B	0.199	0.188	0.182	0.113	0.219	0.228
Qwen3-30B-A3B-Instruct	0.190	0.174	0.134	0.103	0.202	0.216
LLaMA-3.3-70B-Instruct	0.175	0.193	0.115	0.119	0.206	0.157
GPT-4o-mini	0.183	0.190	0.137	0.095	0.181	0.185
GPT-4o	0.165	0.153	0.116	0.092	0.148	0.159
Claude-Opus-4	0.127	0.136	0.123	0.072	0.126	0.148

Table 3.4: Mean sensitivity across models by factor—language (EN vs. FA), input order (Original vs. reverse), and prompt style (Elaborate vs. Concise). Higher values indicate greater prediction volatility; sensitivities are averaged over four prompt variants within each experimental condition (prompt language and input ordering).

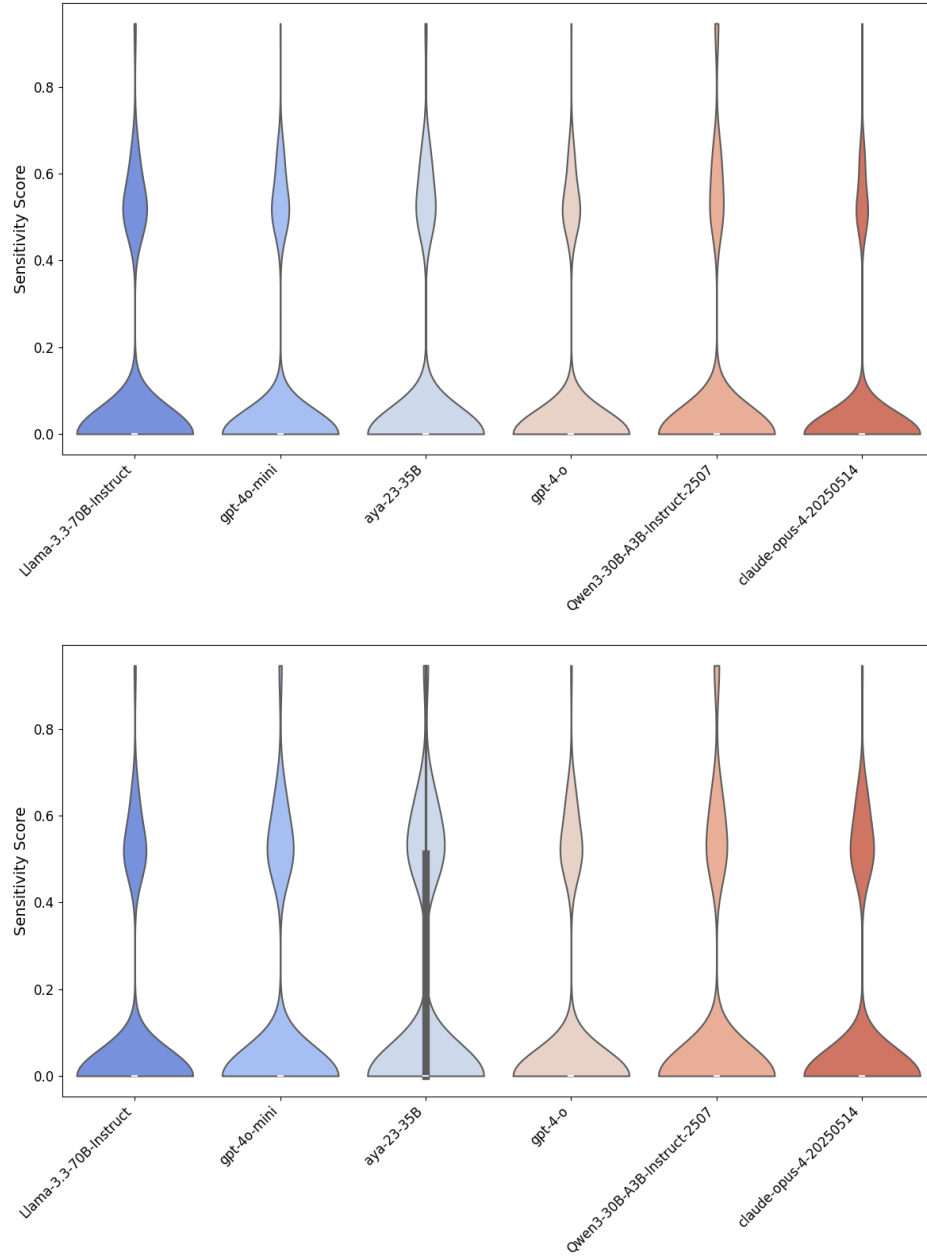


Figure 3.4: Comparison of model sensitivity distributions under different prompt orders. The upper plot shows sensitivity scores when the premise precedes the hypothesis (original order), while the lower plot reverses the order. Across all models, reversing the input sequence increases overall sensitivity, particularly for open-weight models such as LLaMA-3.3-70B and Aya-23-35B, indicating that the order of the premise and hypothesis substantially influences prediction stability.

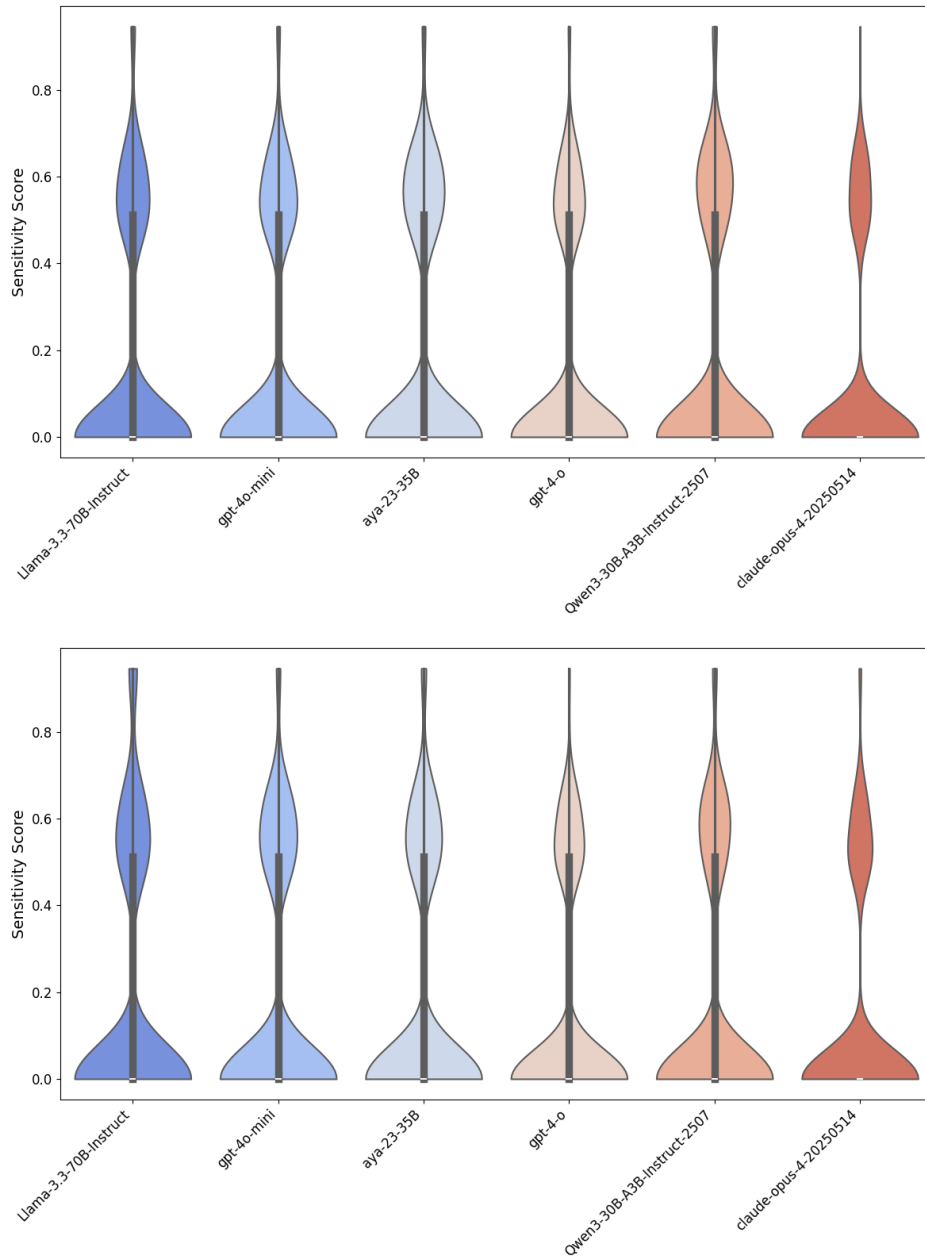


Figure 3.5: Comparison of model sensitivity distributions under different system prompt languages (English vs. Farsi). The upper plot shows sensitivity scores for the English system prompt, and the lower plot shows sensitivity scores for the Farsi system prompt.

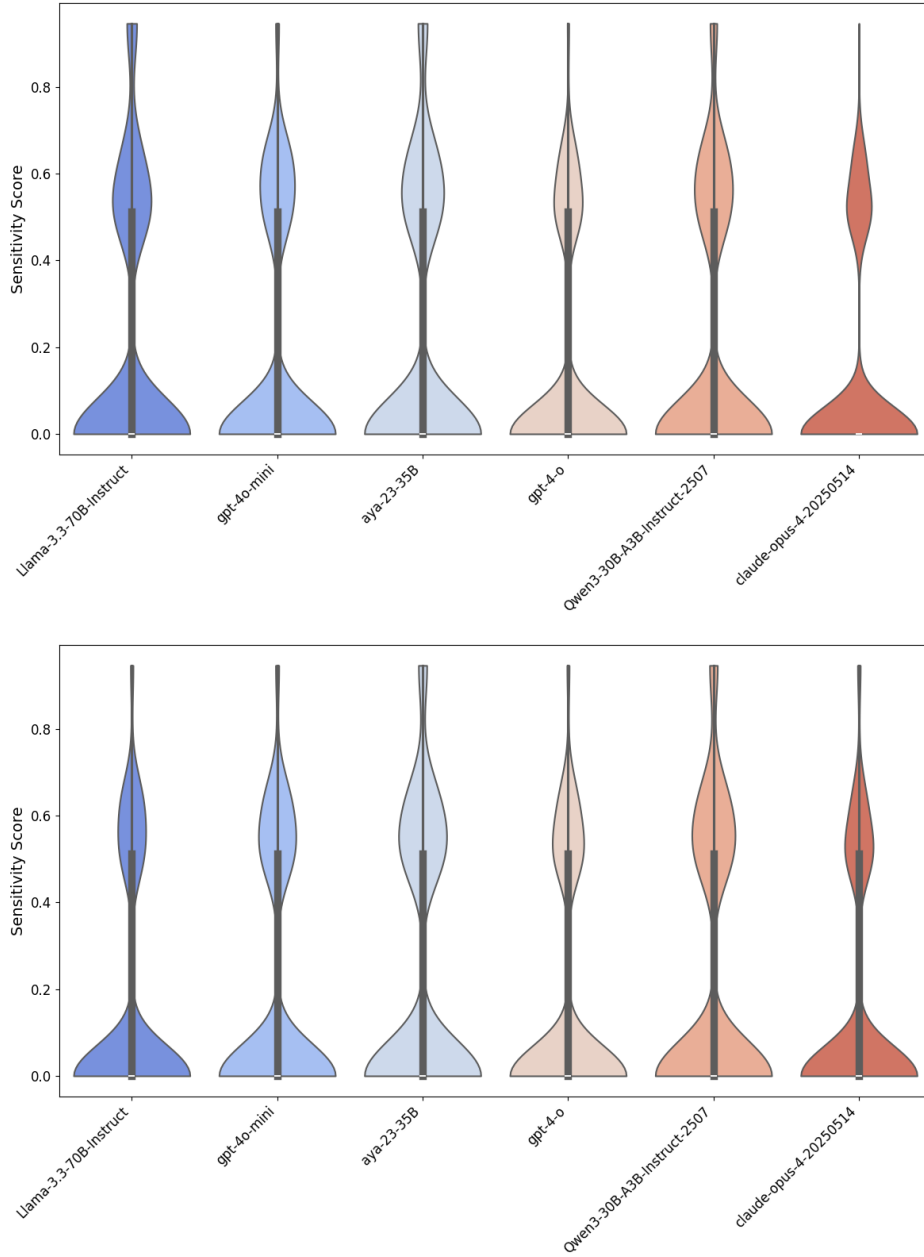


Figure 3.6: Comparison of model sensitivity distributions under different system prompt styles. The upper plot shows sensitivity scores when the system prompt is elaborated, and the lower plot shows sensitivities when the system prompt is concise.

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. Figure 3.7 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 Qwen3-30B exhibit a higher proportion of consistently incorrect predictions, indicating that stability alone does not guarantee correctness. While the underlying causes of this behavior—such as differences in training data or model design—are not directly examined in this study, the results show that prediction stability alone is not sufficient to assess annotation reliability.

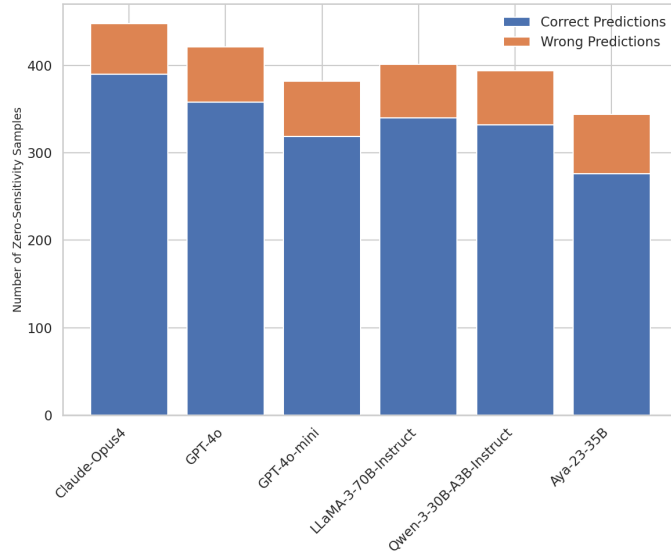


Figure 3.7: Correct and incorrect predictions among zero-sensitivity samples for each model (total dataset: 750 samples). Each bar represents the total number of samples for which a model produced identical labels across all prompt variants (i.e., zero-sensitivity cases). The blue segment indicates predictions that align with the gold-standard labels, while the orange segment denotes consistent but incorrect predictions. Proprietary models such as Claude-Opus-4 and GPT-4o achieve higher overall stability with fewer confident errors, whereas open-weight multilingual models like Aya-23-35B and Qwen3-30B exhibit a greater proportion of consistently incorrect predictions, suggesting that stability does not always imply correctness.

3.3.3 Consistency (Class-Level)

Intra-class consistency assesses the extent to which a model assigns uniform predictions to different inputs belonging to the same ground-truth class across prompt variants [29]. Formally, let $\mathcal{D}_c = \{x_1, x_2, \dots, x_n\}$ be the subset of inputs labeled with class c in the gold dataset, and let $P_{x_i} \in \mathbb{R}^K$ denote the empirical distribution over K output classes predicted for input $x_i \in \mathcal{D}_c$ across multiple prompt variants. The *Total Variation Distance* (TVD) between two such distributions P_{x_i} and P_{x_j} is defined as:

$$\text{TVD}(P_{x_i}, P_{x_j}) = \frac{1}{2} \sum_{k=1}^K |P_{x_i}(k) - P_{x_j}(k)| \quad (3.2)$$

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

$$\text{Consistency}(c) = 1 - \frac{2}{n(n-1)} \sum_{x_i, x_j \in D_c} \text{TVD}(P_{x_i}, P_{x_j}) \quad (3.3)$$

Where $n = |D_c|$ 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. Consider a model evaluated on a class containing three gold-labeled Entailment examples ($n = 3$) across eight prompt variants. Suppose the empirical prediction distributions for each example are as follows:

$$\begin{aligned} P_{x_1} &= (0.75, 0.25, 0.00) \quad [6E, 2N, 0C \text{ out of 8 prompts}] \\ P_{x_2} &= (0.875, 0.125, 0.00) \quad [7E, 1N, 0C \text{ out of 8 prompts}] \\ P_{x_3} &= (0.25, 0.375, 0.375) \quad [2E, 3N, 3C \text{ out of 8 prompts}] \end{aligned}$$

Examples x_1 and x_2 behave similarly because the model predominantly predicts Entailment for both, with minor Neutral leakage. Example x_3 , however, shows a very different profile: despite sharing the same gold label, the model spreads its predictions roughly evenly across all three classes. The pairwise TVD values capture this divergence:

$$\begin{aligned} \text{TVD}(P_{x_1}, P_{x_2}) &= \frac{1}{2} (|0.75 - 0.875| + |0.25 - 0.125| + |0 - 0|) = 0.125 \\ \text{TVD}(P_{x_1}, P_{x_3}) &= \frac{1}{2} (|0.75 - 0.25| + |0.25 - 0.375| + |0 - 0.375|) = 0.5 \\ \text{TVD}(P_{x_2}, P_{x_3}) &= \frac{1}{2} (|0.875 - 0.25| + |0.125 - 0.375| + |0 - 0.375|) = 0.625 \end{aligned}$$

The mean TVD across all $\binom{3}{2} = 3$ pairs is $(0.125 + 0.5 + 0.625) / 3 = 0.417$, giving:

$$\text{Consistency}(\text{Entailment}) = 1 - 0.417 = 0.583$$

This moderate consistency score reflects the fact that x_3 behaves very differently from x_1 and x_2 under prompt variation, even though all three share the same gold label. A consistency of 1.0 would indicate that every Entailment example elicits an identical prediction distribution across prompts; a value near 0 would indicate that the model treats same-class examples as though they belonged to entirely different categories. In

practice, low consistency for a given class suggests that the model lacks a stable internal representation of that class boundary—a pattern we observe most strongly for the Neutral class across all models.

For unambiguous inputs, a consistent LLM is expected to produce similar output distributions across minor variations in the prompt. In such cases, consistency reflects the model’s capacity to generalize within a class under superficial input perturbations. However, for inherently ambiguous premise–hypothesis pairs, some variability may reflect legitimate uncertainty rather than a model deficiency. Importantly, consistency can remain high even when a model is sensitive to prompt variation; in such cases, high consistency indicates systematic rather than random errors, in which 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 3.5 presents class-level consistency scores, which capture a dimension of model behavior that per-class accuracy alone cannot reveal. While the accuracy analysis in Figure 3.2 established that Neutral is the most difficult class across all models, the consistency analysis reveals a complementary pattern: models differ not only in *how often* they are correct on each class, but also in *how stable* their predictions are within a class. The proprietary models (Claude-Opus-4 and GPT-4o) exhibit relatively even consistency across the three classes, with no class falling far below the others (range ≈ 0.13). Their predictions for any given gold class are similar across prompt variants, regardless of whether they are correct. The open-weight models (Aya-23-35B, Qwen3-30B-A3B, LLaMA-3.3-70B), by contrast, show a much wider spread between their best and worst classes (range $\approx 0.30 - 0.35$): they are highly stable on Entailment but unstable on Neutral. This indicates that the open-weight models converge on the same prediction for Entailment examples regardless of how the prompt is phrased, but diverge under prompt variation when the input is Neutral.

Model	Entailment	Contradiction	Neutral
Aya-23-35B	0.83	0.62	0.48
Qwen3-30B-A3B-Instruct	0.76	0.712	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 3.5: Class-wise consistency scores (1 - TVD) for each model, computed across prompt variants. Higher values indicate greater intra-class predictive uniformity.

The consistently low scores for the Neutral class across most models suggest systematic confusion patterns. To further analyze inter-class confusion, we identified two-class confusion samples by selecting instances in which model predictions across all prompt variations were split exclusively between two labels, with zero predictions for the third label. This strict criterion ensured that we examined genuine binary ambiguities in which models demonstrated clear confusion between two specific categories while completely excluding the third. Figure 3.8 shows that all models exhibit the highest confusion between Entailment and Neutral (E–N), indicating that fine-grained semantic distinctions between partial and full Entailment remain a persistent challenge. In contrast, Contradiction–Entailment (C–E) confusion is considerably lower, suggesting clearer separation between logically opposed relations. Meanwhile, Contradiction–Neutral (C–N) confusions vary substantially across models. These findings confirm that most inconsistencies arise from distinctions involving the Neutral class, highlighting a common weakness among models in handling pragmatic relationships between sentences.

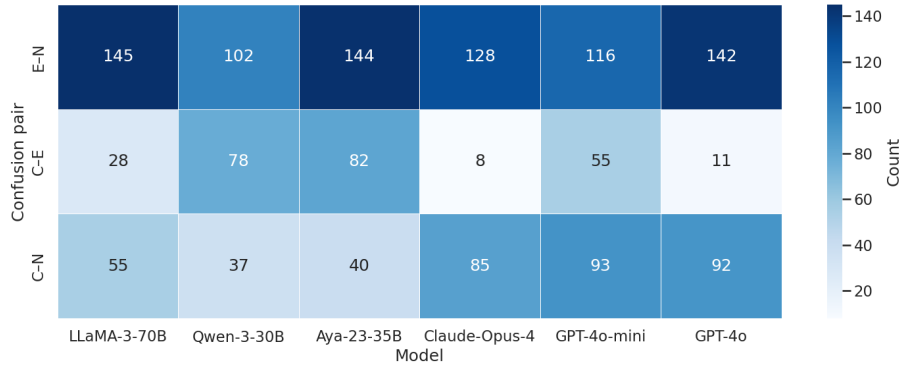


Figure 3.8: Two-Class Confusion frequencies across models (total dataset: 750 samples). This chart displays the count of instances where model predictions were confused between pairs of classes (Entailment-Neutral, Contradiction-Entailment, Contradiction-Neutral) across different prompt variations. The results confirm that Entailment-Neutral is the most prevalent confusion, indicating a shared challenge in distinguishing these semantically subtle classes.

3.3.4 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 in data quality. While NLI appears to demand objective logical reasoning, 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 inputs for which the model’s predicted label changes frequently across prompt

variations. We focused on instances in which 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. 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. The following pages provide examples 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.

GPT-4o-mini

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.

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. Therefore, the relationship is classified as Neutral.

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.

GPT-4o

Premise: They ate in silence and enjoyed the cool shade of a tree. **Hypothesis:** They were standing under a tall oak tree and silently enjoyed the sunlight. **Gold Label:** Contradiction.

Prompt: original_elaborate_en

CoT Reasoning: *Contradiction*. The premise states that they ate in silence and enjoyed the cool shade of a tree, which implies they were sitting or resting under the tree. The hypothesis, however, states that they were standing under a tall oak tree and enjoying the sunlight quietly. These two scenarios are contradictory because the premise involves enjoying the shade, while the hypothesis involves enjoying the sunlight. Additionally, the premise does not specify the type of tree, whereas the hypothesis specifies an oak tree.

Prompt: reverse_concise_en

CoT reasoning: *Neutral*. The hypothesis states that "they stood under a tall oak tree and quietly enjoyed the sunlight," while the premise states that "they ate in silence and enjoyed the cool shade of the tree." The two statements describe different activities (standing vs. eating) and different experiences (enjoying sunlight vs. enjoying shade). There is no direct Contradiction between the two, as both can occur at different times or under different circumstances. However, the premise does not provide enough information to definitively entail the hypothesis, as it does not mention standing or sunlight. Therefore, the relationship is Neutral.

Claude-Opus-4

Example 1:

Premise: The end user is considered a member in the learning process and is not merely a receiver of information. **Hypothesis:** The end user receives information.

Gold Label: Entailment.

Prompt: reverse_concise_en

CoT reasoning: *Contradiction*. The relationship between these two statements is **Contradiction**. The premise states: "The end user is considered a member in the learning process and is not merely a receiver of information." The hypothesis states: "The end user receives information." The key Contradiction lies in how the end user's role is characterized: 1. The Hypothesis presents the end user simply as someone who "receives information." 2. The premise explicitly states that the end user is "NOT merely a receiver of information" and emphasizes their active role as "a member in the learning process". While the premise doesn't deny that the end user receives information at all, it directly contradicts the limited characterization presented in the hypothesis.

Prompt: original_elaborate_en

CoT Reasoning: *Entailment*. Let me analyze the premise and hypothesis: Premise: "The end user is considered a member in the learning process and is not merely a receiver of information." Hypothesis: "The end user receives information." The premise states that the end user is NOT MERELY a receiver of information, but is considered a member in the learning process. This implies that while the end user does more than just receive information (they actively participate in the learning process), they still DO receive information as part of their role. The word "merely" is key here—it doesn't deny that the end user receives information, but rather emphasizes that this is not their only function. Therefore, the hypothesis that "the end user receives information" can be inferred from the premise, as the premise acknowledges information reception as one component of the end user's role, while emphasizing it's not the only component. Answer: Entailment

Table 3.6 shows examples where none of the models predicted the gold label. Models systematically diverge from the original annotations, suggesting either inherent ambiguity in the examples or errors in the gold labels. For instance, in the first example, all models lean toward Entailment or Neutral despite the gold label being Contradiction, implying that the connection between receiving a prestigious medal and the organizational

3.3. RESULTS

excellence of the Iranian Air Force may be inferable rather than strictly contradictory. Similarly, in the fourth example, every model predicts Contradiction, whereas the gold label is Neutral, highlighting a discrepancy in how models and human annotators interpret intentionality and evidence. These cases underscore the critical influence of annotation subjectivity and contextual reasoning in NLI tasks.

Premise (EN)	Hypothesis (EN)	Gold Label	Entailment	Neutral	Contradiction
He was one of the designers and masterminds behind the most complex aviation operation in the world’s military history – Operation H3 – and later received the Fajr (Conquest) Medal from the Iranian Air Force for this achievement.	The Iranian Air Force was one of the best aviation organizations in the Middle East.	Contradiction	13	27	-
It’s as if even if 250,000 people died elsewhere, they don’t want to tell you. You know, they prefer to spend more time working on a 105-year-old cross-shaped shield than on anything else.	A large number of people died elsewhere.	Contradiction	35	5	-
While natural intelligence falls within the 90 to 110 range, the 71 to 84 range is termed Borderline Intellectual Functioning. These individuals are not intellectually disabled but show limited academic achievement.	Borderline intelligence, natural intelligence, and emotional intelligence are distinct types of intelligence.	Contradiction	15	25	-
But such an exhibition would force the museum to take a hard look at its own role in the often-controversial art world.	This museum did not want to examine its role in the art world too closely.	Neutral	-	-	40
There were days when I’d wake up and say to myself: ‘Please God, let me die.	I was suffering from severe depression	Neutral	40	-	-

Table 3.6: Examples where none of the models predicted the gold label.

Finally, building on the earlier sensitivity analyses, we conducted a manual review of instances where the Claude-Opus-4 (the state-of-the-art model in our experiments) exhibited non-zero sensitivity, resulting in 198 samples (26% of the entire dataset). To assess annotation reliability, we compared the predictions on these samples with the human-annotated gold labels. Among these, 94 (47.5%) involved Entailment, 82 (41.4%) Neutral, and 22 (11.1%) Contradiction labels. Manual inspection revealed that 101 (51.0%) were attributable to model misclassification relative to the gold label, while 97 (49.0%) were attributable to issues in the gold dataset. 41 instances were relabeled, and 5 (2.5%) were removed due to low quality or insufficient semantic content. The remaining disagreements reflected borderline distinctions in inference rather than clear annotation errors (Table 3.7).

The comparison between Figure 3.1 and Figure 3.9 reveals a consistent improvement across all evaluated models, with gains of up to several percentage points depending

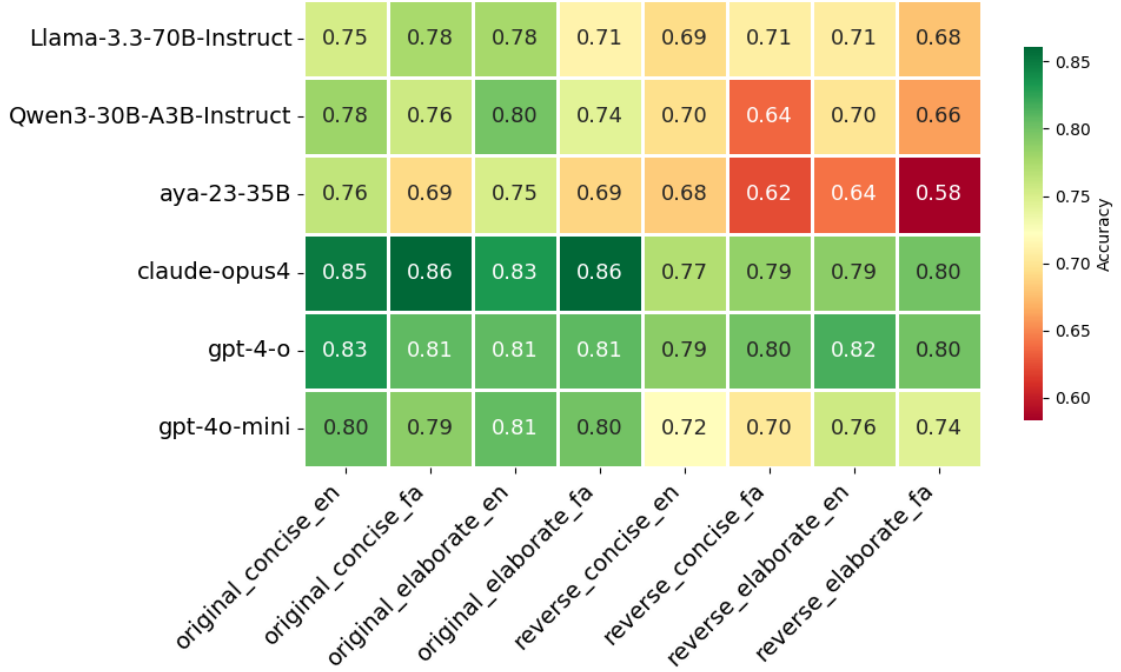


Figure 3.9: Accuracy comparison of six language models across eight prompt configurations for NLI tasks after refining the gold dataset.

on the prompt variant. The improvement is particularly pronounced for closed-source models such as Claude-Opus and GPT-4o, which already demonstrated higher baseline performance. These models benefit more from removing mislabeled and ambiguous cases, suggesting that their predictions were often correct but penalized by unreliable ground-truth annotations. By contrast, open-weight models (Llama-3.3, Aya-23-35B, Qwen3-30B) show smaller relative improvements.

Although this analysis demonstrates that part of the performance gap can be attributed to imperfections in the gold dataset, we deliberately retained the original unmodified dataset for subsequent experiments. This decision ensures that all downstream evaluations, particularly those in Chapter 4 on synthetic data generation and fine-tuning, remain directly comparable to prior benchmarks and faithfully reflect the challenges of working with naturally noisy, human-annotated corpora.

3.3.5 Inter-Model Agreement: Cohen’s Kappa and Fleiss’ Kappa

Beyond individual model behavior, we additionally examine inter-model agreement patterns to assess whether diverse families converge on consistent annotations or exhibit systematic divergences in their interpretive frameworks. To quantify consensus across different model families and prompt configurations, we employ two established statistical

3.3. RESULTS

Premise	Hypothesis	Gold	LLM	Action
<i>Relabeled Examples</i>				
The Byzantines took advantage of the existing instability and advanced their forces beyond the region.	The Iranians were defeated by the Byzantines outside Malatya.	Entailment	Contradiction	Relabeled (Neutral)
Iraqi officials reported that the explosion resulted in significant casualties in central Iraq.	Eight soldiers were killed in an explosion at a munitions depot in central Iraq.	Entailment	Neutral	Relabeled (Contradiction)
<i>Removed Examples</i>				
"So what about me?"	"I feel the same way."	Entailment	Neutral	Removed (Low Context)
Western Malaysia is referred to as Sundaland in geographical studies.	The Malay Peninsula and the islands of Sumatra and Java are located nearby.	Entailment	Neutral	Removed (Ambiguous)
<i>Kept (Ambiguous / Subjective)</i>				
Lincoln stared intently.	The man winked.	Contradiction	Neutral	Kept(Subjective)
Ehsan sat and watched the dust rising in the distance as the riders moved away.	Ehsan watched three riders as they gradually disappeared from view.	Entailment	Neutral	Kept(Borderline)

Table 3.7: Representative examples from the 198 high-sensitivity cases. Relabeled cases reflect corrected annotation inconsistencies; Removed cases lacked sufficient semantic content; Kept cases illustrate borderline inference distinctions.

measures of inter-annotator agreement. Cohen’s Kappa [22] assesses pairwise agreement between two raters, while Fleiss’ Kappa [33] generalizes this framework to multiple raters, both of which account for agreement expected by chance. In our experimental context, these metrics transform our setup into a multi-annotator study, in which each model-prompt combination serves 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. Table 3.8 shows pairwise Cohen’s Kappa scores across all prompt variants, revealing remarkable 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). Figure 3.10 shows detailed per-prompt comparisons.

3.3. RESULTS

	gpt-4o	gpt-4o-mini	Llama3.3	Qwen3-30B	Aya-23	Human
Claude-Opus-4	0.79	0.68	0.66	0.63	0.57	0.72
gpt-4o		0.71	0.64	0.62	0.54	0.69
gpt-4o-mini			0.70	0.73	0.62	0.63
Llama-3.3				0.78	0.73	0.61
Qwen3-30B					0.70	0.59
Aya-23						0.57

Table 3.8: Pairwise Cohen’s Kappa scores across all prompt variants, indicating the degree of agreement between model predictions. Higher values reflect stronger alignment in label choices across the full dataset. Claude and GPT-4o exhibit the highest agreement, whereas Aya-23-35B shows lower agreement with the other models.

Table 3.9 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 [66], 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 models’ best agreement is 0.69. The lower overall agreement when combining all models underscores fundamental behavioral differences across model 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 3.9: 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.

3.4 Discussion

This study set out to answer RQ1: *How reliable are large language models when used as automatic annotators in Farsi NLI?* Our experiments decompose LLM annotation reliability into three interacting dimensions—model family, prompt formulation, and benchmark quality—each contributing distinct failure modes. The key insights 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.

Reliability Gaps Across Model Families. Our results demonstrate a clear reliability hierarchy between model families. Proprietary (Close) models (Claude-Opus-4, GPT-4o) exhibited significantly greater robustness, with roughly 25-30% lower sensitivity (Table 3.3) and better alignment between prediction stability and correctness (Figure 3.7). Several factors plausibly contribute to this gap —instruction-tuning quality, training-data scale and diversity, parameter count, and tokenization—though our experimental design does not disentangle them. 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.

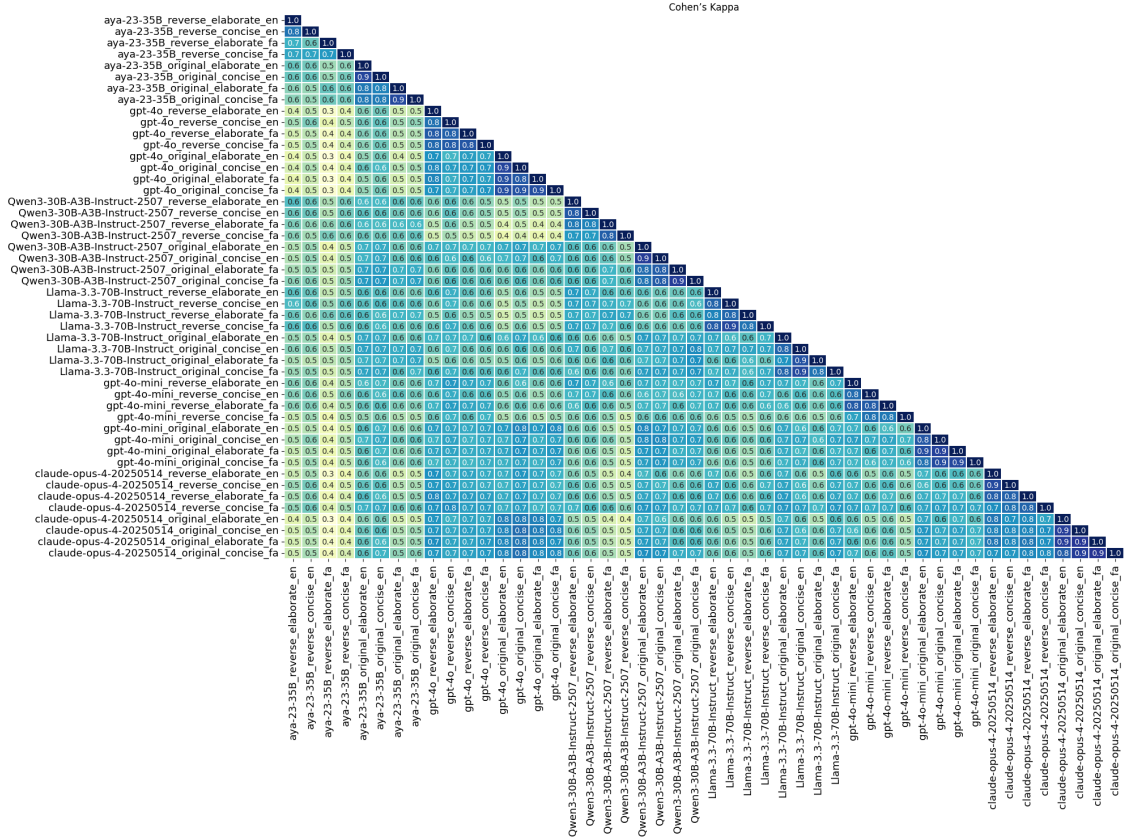


Figure 3.10: Heatmap of pairwise Cohen’s kappa between all model–prompt configurations

Prompt Design as a Key Reliability Factor. The substantial impact of prompt formulation on annotation quality directly addresses the reproducibility concerns in using LLMs as annotators. Notably, we found that ordering (premise-hypothesis sequence) induced greater prediction volatility than language choice or instruction style in terms of sensitivity (Figures 3.4-3.6). However, language choice still influences overall accuracy—as indicated in our grouped analyses by prompt language and style (Figure 3.1)—suggesting that while structural consistency supports reliability, prompt language continues to play a role in final performance. We hypothesize that the reduced labeling performance under reverse ordering stems from training data conventions, particularly from consistent exposure to sequences in which the premise precedes the hypothesis, though we cannot verify this directly given the closed nature of most models’ training data. However, analysis of the models’ reasoning on high-sensitivity samples reveals that LLMs are not relying solely on surface-level heuristics but can access multiple, often independently defensible, reasoning paths. Our chain-of-thought analysis demonstrates that the same model, when given different prompts, can articulate distinct yet logically sound interpretations of the

same premise-hypothesis pair. Some of the prompt-induced variability we observe may reflect genuine ambiguity in the underlying examples—the same kind of ambiguity that produces inter-annotator disagreement among humans. We do not test this hypothesis directly here; a systematic comparison of LLM and human disagreement patterns on identical items is left to future work. However, we emphasize that not all prompt-induced variation reflects legitimate interpretive flexibility; a substantial share likely reflects genuine instability. Empirically distinguishing the two remains an open problem. The fact that models can be steered towards these different valid interpretations through prompt design suggests that some of what appears as instability may instead reflect a form of contextual reasoning flexibility—though, as noted above, this cannot account for all observed variability.

The Neutral Class as a Reliability Challenge. Across all models and metrics, the Neutral class exhibits the lowest consistency scores (Table 3.5) and highest confusion rates (Figure 3.8), indicating that distinguishing partial from full Entailment remains the central open challenge for LLM-based NLI annotation. 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 in section 3.3.4 revealed that approximately 49% of high-sensitivity cases involved noisy gold labels or inherent ambiguities. After cleaning the dataset, we observed consistent performance improvements across all models (Figure 3.9), with proprietary models showing the largest gains, averaging 4-5 percentage points. This finding crucially reframes our understanding of reliability: what appears as model instability may sometimes reflect legitimate disagreement with problematic annotations. They also reinforce findings from prior works, which show that human annotation disagreement is not merely noise but often reflects systematic and valid differences in interpretation. For instance, Plank [102] argues that many NLP datasets rely on an oversimplified notion of a single “ground truth” label, despite the inherent ambiguity of language. Similarly, Jiang et al. [60] demonstrate that label variation can be predictable and correlated with annotator perspectives, rather than random error. In this context, the improvements observed after dataset cleaning suggest that some disagreements in the original data may stem from inconsistent or low-quality annotations, rather than genuine ambiguity. However, these findings also caution that not all disagreement should be treated as noise, as it may reflect a distribution of valid interpretations rather than a single correct label. This observation highlights the importance of dataset quality in low-resource settings, where the scarcity of qualified annotators often leads to noisy or inconsistent labels. As a result, benchmark performance may re-

flect annotation artifacts rather than the model’s true capability, calling for more careful, nuanced evaluation.

3.5 Summary

In this chapter, we presented a comprehensive analysis of the reliability of LLMs as annotators for the NLI task in Farsi. The findings confirm the critical dependence of annotation reliability on prompt formulation, model family, and instruction language. Overall, proprietary models are more robust; open models are less stable. Reliability is strongly shaped by prompt design, especially input ordering and instruction language. The Neutral class remains the main challenge, and much observed instability is driven by dataset noise rather than true model failure. These results emphasize that reliable LLM-based annotation requires not only careful model and prompt selection, but also critical attention to benchmark quality and evaluation practices. In the following chapters, we introduce our pipeline and examine how variations in LLM behavior, as influenced by annotators, affect the downstream task.

4

LLM-Driven Data Generation and Labeling at Scale

4.1 Introduction

In this chapter, we present a scalable LLM-based framework that unifies data generation and weak annotation into a single pipeline, enabling efficient creation of synthetic datasets for low-resource languages and mitigating data scarcity. LLMs are capable of producing examples that closely mimic human-written data—often with substantial cost and time savings. In addition, generating data with LLMs can yield more diverse samples due to their probabilistic nature. Therefore, linguistic variations in such a dataset are not limited by the coverage of hand-designed rules. Natural Language Understanding (NLU) tasks, such as NLI, require deliberately constructed non-Entailment examples. Prior work has relied on rule-based perturbation strategies to generate such examples: for instance, Kryscinski et al. [68] applied sentence negation, entity swapping, and pronoun replacement to create factually inconsistent summaries for training consistency classifiers, while Glockner et al. [41] introduced lexical substitution-based perturbations to probe NLI model robustness. Similarly, the ANLI dataset [93] employed human-in-the-loop adversarial annotation, in which annotators manually craft hypotheses designed to fool a target model—a process that, while producing high-quality challenge examples, is inherently time-consuming and limited in scale. These perturbation-based and manual annotation approaches share a common limitation: they rely on hand-crafted transformation rules or costly human effort, constraining both the diversity of the resulting negative examples and the scalability of dataset construction.

LLMs have recently been leveraged to address this challenge by automatically generating such contrastive examples at scale. For instance, the TrueTeacher [37] pipeline utilizes LLMs of varying capacities to produce millions of non-factual summaries (i.e., non-entailing) for training factual inconsistency classifiers. Beyond generating text, several studies have explored LLMs as annotators that provide weak supervision signals in place of costly crowdsourced labeling [40, 51, 28]. Other recent works have examined the

impact of synthetic data on downstream tasks, showing that models trained on LLM-generated data can approach or even surpass those trained on gold-standard annotations, while also offering a scalable alternative to crowd-based labeling [65, 12].

Building on this paradigm, we propose a scalable, cost-effective, LLM-based pipeline that uses large models for both generation and annotation of synthetic NLI datasets. Unlike TrueTeacher, which focuses on NLI as an evaluation mechanism for factual consistency in English summarization, our method applies the framework directly to NLI data creation for Farsi. While our approach is designed to be task- and language-agnostic, we focus on Farsi as a case study to demonstrate its effectiveness in a genuinely low-resource setting. In total, in our experiments, our pipeline produces and labels over 100,000 sentence pairs, providing a large-scale synthetic dataset for examining the reliability and transferability of LLM-generated supervision. We further evaluate the utility of the resulting synthetic data in downstream NLI tasks by training models on the generated dataset and analyzing their performance across different data scales and annotation settings.

The remainder of this chapter is organized as follows. Section 4.2 introduces the proposed LLM-driven pipeline, detailing its generation and annotation components, and Section 4.3 discusses the case study on NLI data generation for Farsi. Section 4.5 describes the experimental setup, including data preparation, models, and evaluation metrics. Section 4.6 presents the empirical results and analysis of model performance across data sizes and annotation sources. Finally, Section 4.7 discusses the broader implications of our findings for synthetic data generation in low-resource settings and outlines potential directions for future research.

4.2 Method

Our proposed framework is designed to generate synthetic data for any text classification tasks involving two input texts, such as textual Entailment, summary verification, question-answer verification, translation quality estimation, etc. An overview of the proposed pipeline is shown in Figure 4.1. Our approach is conceptually inspired by TrueTeacher [37], which used large language models as weak teachers to generate factual-consistency labels for English summaries. In contrast, we extend this teacher-based paradigm beyond factuality, applying it to large-scale synthetic data generation for classification tasks in low-resource languages.

First, we use an LLM as a generator to generate pairs of texts D_{gen} , where each pair consists of a preceding text (x_p) and a succeeding text (x_s). In the NLI task, x_p will be a premise and x_s its candidate hypothesis; in summary verification, x_p will be a document and x_s its candidate summary. In our framework, a generator LLM creates candidate pairs by selecting good-quality source texts for (x_p) and generating appropriate

4.3. CASE STUDY: DATA GENERATION FOR FARSI NLI

successors (x_s) for the task at hand. Then by using an LLM as a judge, each pair in D_{gen} will be annotated with the help of task-specific judgment instructions and generate a labeled dataset D_{lab} , where,

$$D_{lab} = \{(x_p, x_s, y) \mid (x_p, x_s) \sim D_{gen}, y \sim p(y \mid x_p, x_s)\}$$

An annotator LLM generates the label y for a given pair of input texts $x = (x_p, x_s)$ according to the judgment instruction.

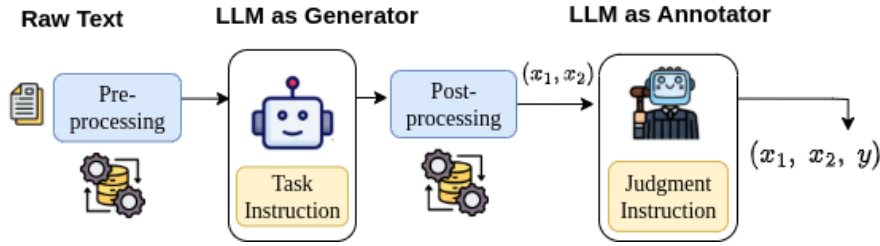


Figure 4.1: Synthetic data generation pipeline for classification tasks with two textual inputs. An LLM, guided by task instructions, generates input pairs (x_1, x_2) from raw text. A second LLM, prompted with judgment instructions, labels the pairs, producing annotated triplets (x_1, x_2, y) for training.

The proposed pipeline is designed to be task-agnostic: it produces labeled text pairs based on criteria (Instructions) that can be flexibly adapted to different downstream tasks. Specifically, pairs labeled as entailed correspond to semantically consistent or valid outputs. Whereas generated samples labeled as negative serve as meaningful examples that reflect real-world failure modes (e.g., hallucination, irrelevance). These curated negative samples can support contrastive learning or reward-based training by helping models distinguish valid inferences from invalid inferences. For classification tasks such as NLI, all labels are retained; for generative tasks such as question answering, summarization, or translation, only entailed pairs are used to construct the dataset; and for contrastive or robustness learning, non-entailed examples serve as negative samples.

4.3 Case Study: Data Generation for Farsi NLI

We generated synthetic Farsi data for the NLI task using the process described in Section 4.2. We first generated pairs of premises and hypotheses (D_{gen}) by using an LLM generator. Then we labeled these pairs with the Entailment, Contradiction, or Neutral classes using an LLM annotator and generated D_{lab} , which is used in our NLI experiments.

4.3. CASE STUDY: DATA GENERATION FOR FARSI NLI

4.3.1 Premise and Hypothesis Generation

We begin our NLI data generation by selecting high-quality samples as premises. The premises are drawn from the Farsi portion of the XLSum dataset [48], ensuring that they are human-written and exhibit a natural linguistic distribution. Since the input texts in XLSum are significantly longer than the samples in the gold-standard ParsiNLU dataset, we apply a filtering process to extract shorter samples with a length distribution similar to that of the gold dataset.

We then generate hypotheses by applying an LLM-based summarization model to the selected premises. Here, our fundamental assumption is that the correct summaries of premise samples will be their Entailment hypothesis, and incorrect summaries will be their Contradiction or Neutral hypothesis. For hypothesis generation, we use the mT5-base model, a multilingual variant of the T5 model [109], fine-tuned on Farsi human-annotated summarization data [31]^a. This domain adaptation enhances fluency and stylistic alignment with native Farsi text generation.

While directly prompting a large language model to generate hypotheses for different classes is a straightforward approach, using mT5-base is more computationally efficient and scalable, enabling cost-effective generation of large volumes of data.

Note that we did not select the best-performing summarization model for Farsi because the incorrect summaries it generates are a valuable resource for generating contradictory and Neutral hypotheses. At the same time, a too-weak model would yield low-quality hypotheses that confuse the annotator LLM, increasing label noise and degrading the dataset quality. Our choice, therefore, reflects a trade-off: sufficient model strength for fluent, informative hypotheses, yet with a controlled error rate that still yields diverse non-entailed cases.

To generate summaries, we adopt the standard T5-style prompt formulation, structured as: "summarize:" + {input_text}. We configure generation with a maximum output of 100 tokens and employ beam search (beam size=2) to reduce variability and mitigate hallucination. However, we observed that over 90% of the generated summaries began with a news agency name (e.g., ISNA news wrote:), even when such attributions were absent in the original text, indicating a bias inherited from pre-training on journalistic corpora. To mitigate this issue, we implemented a post-processing step to identify and remove these prefixes while preserving the semantic integrity of the summaries. Additionally, we observed that some of the hypotheses include various forms of noise such as grammatical errors, hallucinated or fabricated content, unattributed paraphrases, and shifts in style or formality. These noisy samples were retained in the dataset to enable future filtering studies and robustness evaluations. We generated more than 100,000

^areported metrics: ROUGE-L: 39.96, BERTScore: 79.54, average generation length: 48.72 tokens.

4.3. CASE STUDY: DATA GENERATION FOR FARSI NLI

premise-hypothesis pairs (D_{gen}), within approximately 6 hours on an A100 (80GB) GPU. Samples of the generated hypotheses are shown in Figure 4.2. The generated summaries naturally exhibit a mixture of Entailment, Contradiction, and Neutral relationships with respect to their corresponding premises.

Premise (FA)	Hypothesis (FA)	Premise (EN)	Hypothesis (EN)
در پیش‌نویس این گزارش فصلی با چهار هزار کلمه به استراتژی نژادپرستانه پلیس نیویورک در برابر مسلمانان اختصاص داده شده بود و این که چه طور پلیس نیویورک بعد از حملات یازدهم سپتامبر برای جوامع مسلمان نیویورک مأمور مخفی و خبرچین گمارده بود.	گزارش فصلی درباره استراتژی نژادپرستانه پلیس نیویورک در برابر مسلمانان به تازگی منتشر شد.	In the draft of this quarterly report, a 4,000-word chapter was devoted to the racist strategy of the New York police toward Muslims, including how, after the September 11 attacks, undercover agents and informants were deployed in Muslim communities.	A quarterly report on the racist strategy of the New York police toward Muslims was recently published.
کشورهای غربی و حتی دولت اسرائیل که بر ادامه تحریم سیاسی و اقتصادی دولت حماس اصرار می‌ورزد، به همان اندازه که خواهان در تنگنا قرار دادن دولت حماسند، از فروپاشی اجتماعی در سرزمینهای فلسطینی نیز بیمناکند.	دولت حماس که به تازگی تحریم‌های اقتصادی و سیاسی علیه رژیم صهیونیستی را اجرا کرده است، از فروپاشی اجتماعی در سرزمینهای لبنانی نیز بیمناکند.	Western countries, and even the Israeli government, which insist on continuing political and economic sanctions against Hamas, while seeking to pressure it, also fear social collapse in the Palestinian territories.	The Hamas government, which has recently imposed political and economic sanctions on Israel, fears social collapse in Lebanese territories.
آیا شما میدانید در زمان هخامنشی بانک بوده و مردم برای هم چک میکشیدند که این کلمه یک کلمه به یادگار مانده از آن دوران است.	در زمان هخامنشی بانک بوده و مردم برای هم چک میکشیدند.	Did you know that during the Achaemenid era there were banks and people used checks, and that this word has remained from that time?	During the Achaemenid era, there were banks and people used checks.
تیم ایران که به دعوت آمریکا برای شرکت در این تورنمنت تابستانی به این کشور سفر کرده است، چند دیدار تدارکاتی برگزار کرده است.	تیم ملی فوتبال ایران برای شرکت در تورنمنت تابستانی در کشور آفریقای جنوبی به دعوت آمریکا وارد این کشور شد.	The Iranian team, invited by the United States to participate in this summer tournament, has traveled there and held several preparatory matches.	The Iranian national football team arrived in South Africa, invited by the United States to participate in the summer tournament.

Figure 4.2: Sample premise–hypothesis pairs generated by the synthetic data pipeline in Farsi, with English translations provided for clarity. The examples illustrate how the summarization-based approach naturally produces Entailment, Contradiction, and Neutral relationships.

Figure 4.3 presents the distribution of sentence lengths—measured by word count—for both premises (left plot) and hypotheses (right plot), comparing the ParsiNLU training, validation, and test splits. The gold splits exhibit consistent and overlapping distributions, indicating a well-stratified sampling process. Figure 4.4 shows that the generated synthetic data (D_{gen}) exhibits a more pronounced right-skewed distribution for both premises and hypotheses, indicating longer and more variable sentence lengths compared to the ParsiNLU training set.

4.3. CASE STUDY: DATA GENERATION FOR FARSI NLI

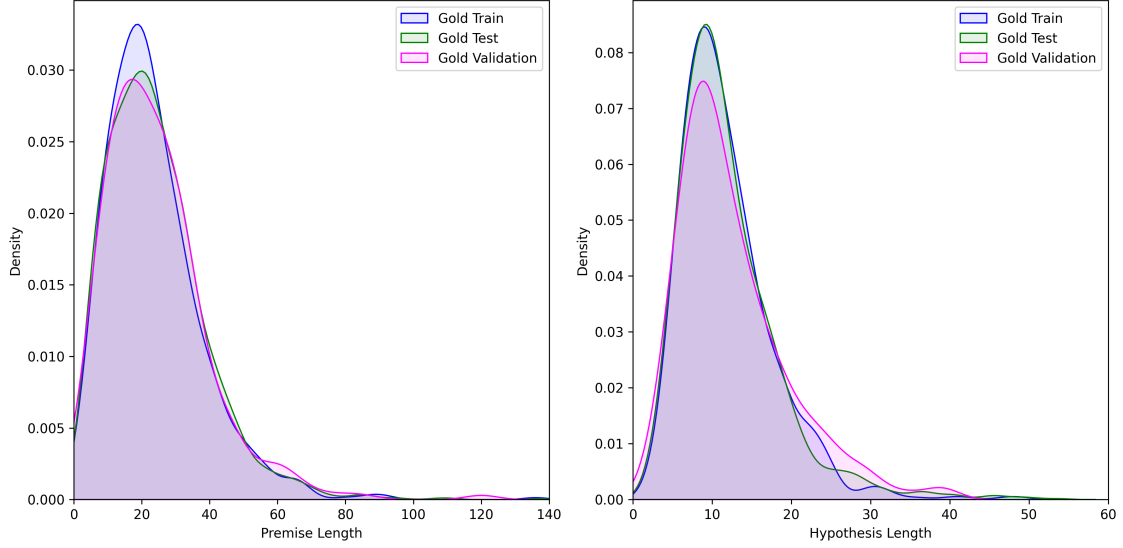


Figure 4.3: Distribution of sentence lengths (in number of words) across gold-standard dataset splits. On average, premise sentences are approximately 23 words long, while hypotheses average around 11 words.

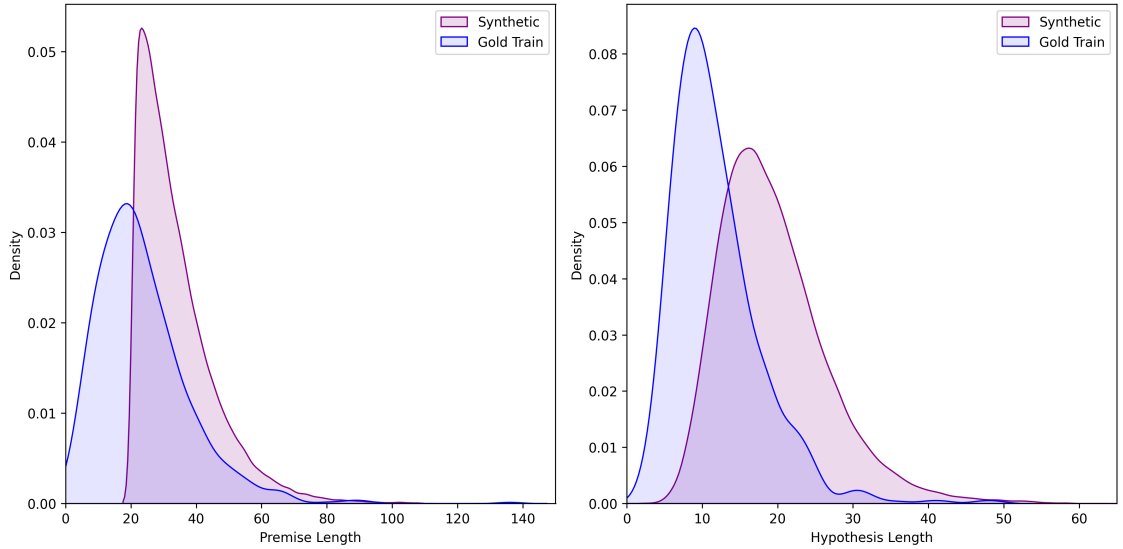


Figure 4.4: Distribution of sentence lengths for premise (left) and hypothesis (right) in the synthetic and gold training datasets. On average, synthetic premise sentences are approximately 33 words long, while synthetic hypotheses average around 19 words.

4.3.2 Labeled Data Generation

Premise-hypothesis pairs generated in the previous section are labeled as Entailment, Contradiction, or Neutral by using a stronger annotator LLM to generate labeled NLI

4.3. CASE STUDY: DATA GENERATION FOR FARSI NLI

synthetic data. We selected one of the best-performing general-purpose language models, GPT-4o-mini, as the annotator LLM for label generation. GPT-4o-mini offers a cost- and latency-efficient option while retaining strong benchmark capability, making it practical for large-scale labeling. Later in Chapter 5, we validate the robustness of our pipeline towards other annotator LLMs by using an open-weight language model, Aya-23-35B, and a small specialized NLI model, DeBERTa-NLI [72]. We use the GPT-4o-mini model with the best prompt obtained (`original_elaborate_en`) from Chapter 3 ^b :

Annotation Prompt

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. Premise:{premise}
Hypothesis:{hypothesis} Answer:

The generation is constrained by choosing only the specified tokens, in this case, the class labels: *Entailment*, *Contradiction*, or *Neutral* ^c. We set the temperature to zero to ensure deterministic outputs and minimize randomness in the model’s responses. Additionally, we limit the maximum number of tokens generated to 5 to enforce concise, single-label answers, ensuring the model strictly adheres to the required categorical response format and avoids unnecessary text.

For the annotation of synthetic data, each labeling request consists of a prompt with approximately 250 tokens, and these prompts were sent in bulk to OpenAI’s Batch API, which allows requests to be sent asynchronously in batches at 50% lower cost (0.15 USD per 1 million tokens)^d. Requests submitted to OpenAI were fulfilled within 24 hours.

We generated over 100,000 labeled data samples, comprising approximately 65,000 Entailment, 32,000 Neutral, and 18,000 Contradiction instances. The label distributions of the synthetic data and the ParsiNLU dataset are shown in Table 4.1. Figure 4.5 visualizes how the label distributions in the synthetic dataset changes as the dataset size increases. The purpose of this analysis is to characterize the intrinsic bias of the generation process: as the dataset grows, the proportion of Entailment, Neutral, and Contradiction labels stabilizes, indicating a persistent skew toward Entailment that does not self-correct with scale. This bias is not an artifact of small-sample variance but a property of the mT5-based generation pipeline itself. This bias reflects that the mT5-base model tends to

^bOther prompting strategies such as few-shot and chain-of-thought (CoT) were not used to maintain consistency across all experiments in this thesis.

^cConstraint sampling is a technique to filter the logits vector to keep only the tokens that meet the constraints [57].

^d<https://openai.com/api/pricing/>

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

produce hypotheses that preserve the original meaning, resulting in a higher proportion of entailed pairs. While using a smaller model, such as mT5-small, could generate more non-entailed samples, we observed that the teacher model struggles to assign accurate Neutral and Contradiction labels when the hypothesis is noisy. In this case, the labels are more likely to be noisy, which degrades the datasets’ overall quality. However, high Entailment rates remain advantageous for other downstream tasks—where only entailed pairs are required.

	Entailment	Neutral	Contradiction
GPT-4o-mini	65492	32865	18331
ParsiNLU (train)	275	241	234
ParsiNLU (validation)	104	87	78
ParsiNLU (test)	608	559	501

Table 4.1: Label distribution across synthetic and gold datasets.

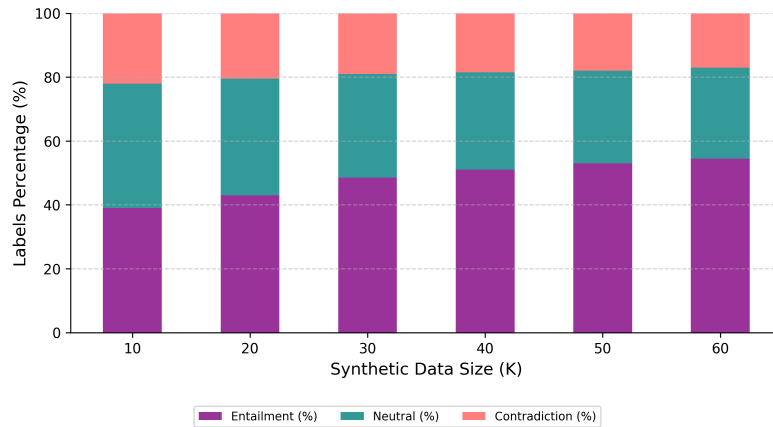


Figure 4.5: Label distribution (%) across different synthetic dataset sizes. The distribution remains skewed toward Entailment as dataset size increases, reflecting a consistent bias in the LLM-based summarization and labeling pipeline.

4.4 Lexical Analysis of Synthetic and Gold Data

Table 4.2 reports vocabulary statistics for the synthetic datasets at each scale alongside the gold dataset. Unique token counts grow steadily with dataset size (22K at 5K examples to 87K at 50K), confirming that additional synthetic examples introduce genuinely new vocabulary. Gold vocabulary coverage—the proportion of gold vocabulary present in the synthetic data—grows from 62.6% at 5K to 79.2% at 50K, but with diminishing returns beyond 20K. The Type-Token Ratio (TTR) of the synthetic data decreases

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

monotonically with dataset size (0.135 at 5K to 0.053 at 50K), compared to 0.322 for the gold premises. This gap is primarily a corpus-size artifact: with up to 155× more examples, tokens naturally repeat more frequently in larger corpora. To isolate the size effect, we computed TTR on a random size-matched subsample of 749 synthetic premises, obtaining a TTR of 0.27—moderately below the gold TTR of 0.322. This modest gap reflects the synthetic data’s reliance on a single journalistic source corpus (XLSum/BBC Farsi), which introduces a degree of stylistic homogeneity compared to the manually curated gold dataset. Despite this, the consistently high gold vocabulary coverage across all scales confirms that the synthetic data remains broadly aligned with the target lexical distribution.

Size	TTR		Unique tokens		Gold coverage	
	Prem.	Hyp.	Prem.	Hyp.	Prem.	Hyp.
5K	0.135	0.102	22,527	9,732	0.626	0.618
10K	0.103	0.073	34,324	13,999	0.686	0.677
20K	0.077	0.052	51,657	20,067	0.738	0.733
30K	0.065	0.043	65,372	24,765	0.764	0.755
50K	0.053	0.034	87,674	32,306	0.792	0.782
Gold (749)	0.322	0.369	5,670	3,274	—	—

Table 4.2: Lexical statistics across synthetic dataset sizes and gold data. TTR decreases with size as a natural corpus-size effect. Unique tokens grow steadily, confirming vocabulary expansion at each scale. Gold coverage reports the proportion of gold vocabulary present in the synthetic data.

Figure 4.6 shows per-example vocabulary overlap ratios between premises and hypotheses, each dataset split, grouped by NLI class. Across all three classes, synthetic distributions mainly match gold train, validation, and test distributions in both shape and median, confirming in-domain lexical consistency. Entailment pairs show greater overlap than Neutral and Contradiction pairs across all splits, consistent with the semantic proximity inherent to entailing premise-hypothesis pairs.

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

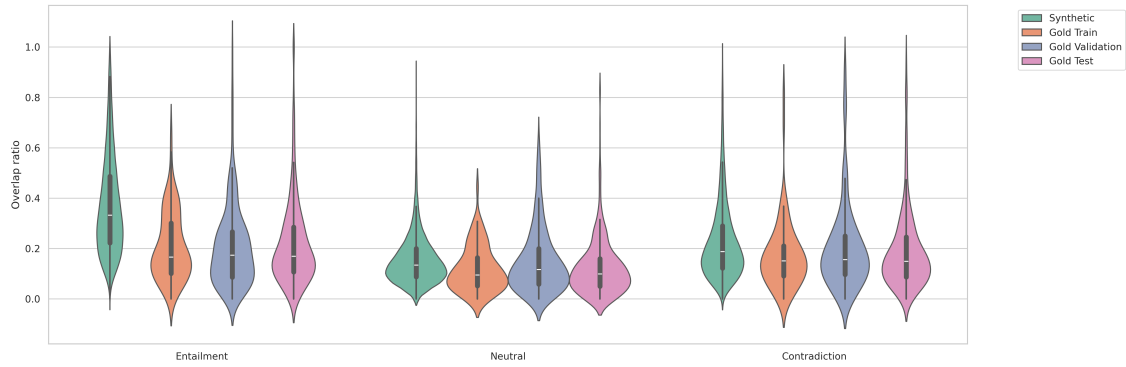


Figure 4.6: Lexical overlap ratios between premise and hypothesis across data sources. Synthetic data shows higher overlap, especially for Entailment, indicating stronger lexical similarity induced by summarization.

Figures 4.7 and 4.8 compare the textual properties of the synthetic and gold datasets across NLI classes, focusing on the length of premise and hypothesis. Synthetic premises and hypotheses are longer on average than their gold counterparts, with the difference most pronounced for Entailment pairs.

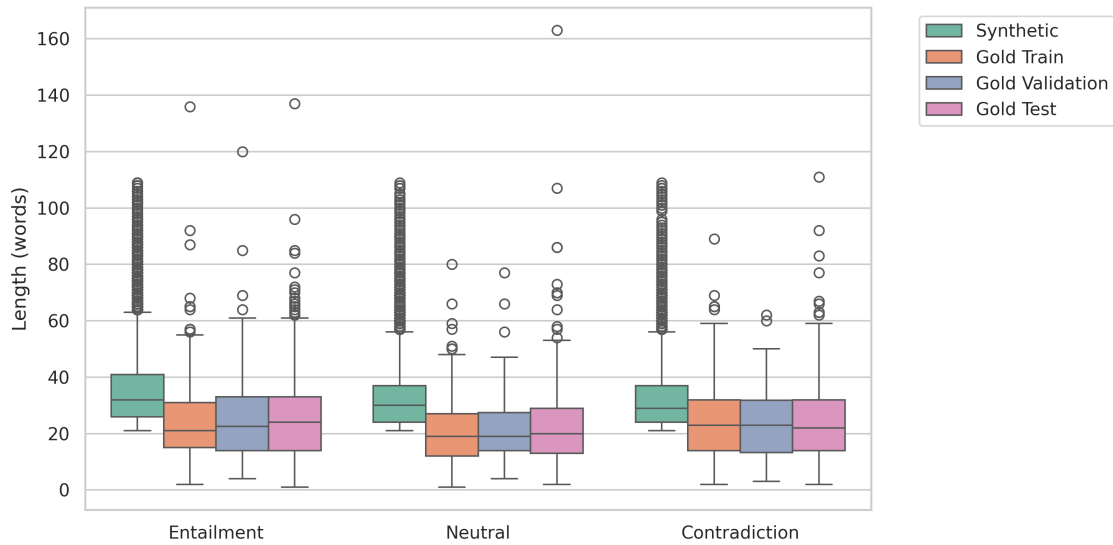


Figure 4.7: Premise length distributions across synthetic and gold datasets, grouped by label. Synthetic premises are generally longer, particularly in the Entailment class.

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

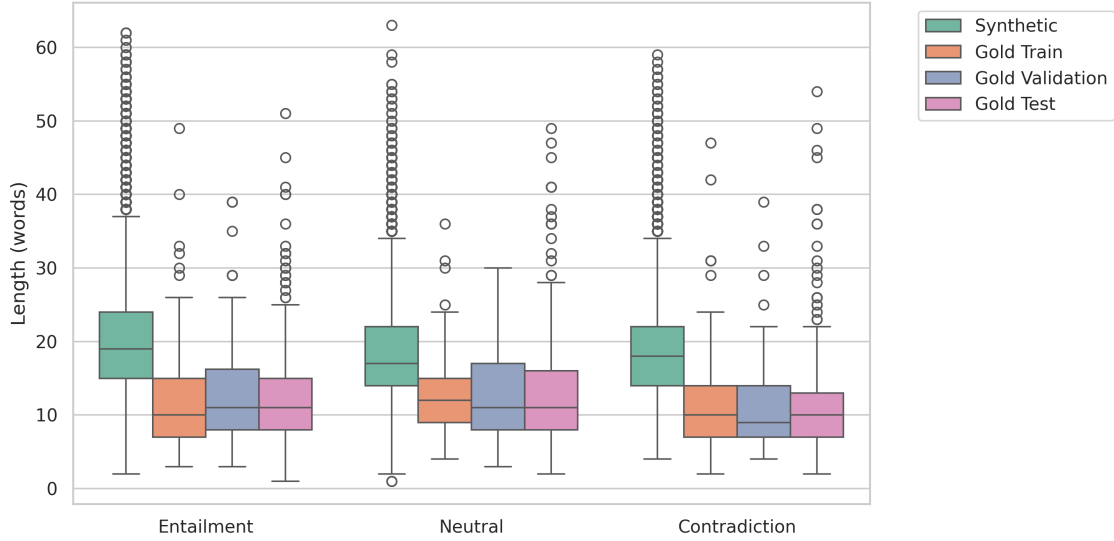


Figure 4.8: Hypothesis length distributions for synthetic and gold data. Synthetic hypotheses are longer overall, consistent with their origin as LLM-generated summaries of the premise.

4.4.1 Human Evaluation on Synthetic Data Quality

We evaluate GPT-4o-mini annotation quality by comparing its labels against human annotations on 600 synthetic sentence pairs. The overall agreement between GPT and human labels is 72.6%, with 436 matched cases and 164 mismatches. To guide the human annotation process, two annotators were explicitly instructed to evaluate Entailment solely based on the textual content of the premise and hypothesis, without drawing on commonsense reasoning or external world knowledge. Specifically, annotators were asked to apply an *according to* test: given only the premise, does the hypothesis follow? This instruction mirrors the attribution principle formalized in the AIS framework [111], which defines a valid inference as one a generic reader would affirm when asked *according to the source, does this follow?*— explicitly excluding background knowledge or real-world assumptions beyond the text. Some of the samples on which LLM and humans disagree are shown below. Manual inspection of disagreement cases reveals four recurring error patterns:

- **Referential ambiguity:** GPT over-confidently resolves underspecified references, such as ambiguous pronouns (“he”) or unnamed locations (“Isfahan”), leading to unwarranted Entailment predictions.
- **World knowledge interference:** The model occasionally draws on external knowledge not present in the premise, producing labels that reflect plausible inferences

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

rather than textual Entailment.

- **Entailment fuzziness:** Subtle lexical substitutions blur the boundary between Neutral and Contradiction, leading to borderline misjudgments.
- **Agency misattribution:** Confusion over speaker or actor identity causes incorrect Entailment assignments when the hypothesis shifts attribution relative to the premise.

Synthetic Data

Premise: During the recent economic crisis, he asked the European Commission to take more decisive action, intervene more in controlling member states' economies, and at the same time guide countries toward closer integration.

Hypothesis: The President of the European Commission demanded greater intervention in controlling member states' economies.

Human: Neutral

GPT-4o-mini: Entailment→Reasoning: The premise states that "he asked the European Commission to take more decisive action" and "intervene more in controlling member states' economies." This implies that the President of the European Commission is advocating for greater intervention, which aligns with the hypothesis that he demanded greater intervention in controlling member states' economies. Therefore, the hypothesis can be inferred from the premise.

Remark: There's ambiguity about who "he" is, the hypothesis is plausible, but not entailed or contradicted.

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

Synthetic Data

Premise: City officials have stated that railway services will gradually return to normal starting early Wednesday morning, and are expected to be operating at an acceptable level by the end of the day.

Hypothesis: Isfahan railway services will return to normal starting early Wednesday morning.

Human: Neutral

GPT-4o-mini: Entailment→Reasoning: The premise mentions that "railway services will gradually return to normal" but does not specify which city's railway services are being referred to. While it is possible that Isfahan is included, the information provided does not confirm it, making the hypothesis uncertain.

Remark: The location-specific detail (Isfahan) is not entailed by the premise.

Synthetic Data

Premise: Jared Kushner said that around \$50 billion would be invested in the region as part of this plan, creating many job opportunities in the West Bank and Gaza.

Hypothesis: Venezuela's Minister of Energy announced that the country intends to invest \$50 billion in the West Bank and Gaza to create many job opportunities.

Human: Neutral

GPT-4o-mini: Contradiction→Reasoning: The premise states that Jared Kushner mentioned a \$50 billion investment in the West Bank and Gaza, but it does not indicate that this investment is coming from Venezuela's Minister of Energy. The hypothesis claims that Venezuela is making this investment, which contradicts the information provided in the premise.

Remark: The model treats differences in named entities or specific details as evidence of Contradiction, even when those details are not mutually exclusive.

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

Synthetic Data

Premise: When they say John Kerry’s son-in-law, the U.S. Secretary of State, is of Iranian descent, this cannot be without influence on his view of Iran and his understanding of our country.

Hypothesis: The U.S. Secretary of State states that when they say he is of Iranian descent, this cannot be without influence on his view of Iran and his understanding of our country.

Human: Contradiction

GPT-4o-mini: Entailment→ Reasoning: The premise discusses the influence of the U.S. Secretary of State’s Iranian descent on his views, but it does not state that he himself makes this claim. The hypothesis suggests that the Secretary of State explicitly states this influence, which is not confirmed or denied by the premise. Therefore, we cannot definitively conclude that the hypothesis is true or false based on the premise alone.

Remark: There is misattribution of identity and speaker.

These disagreements are consistent with challenges reported in human AIS annotation studies, where annotators similarly struggle with referential ambiguity and named entity underspecification despite explicit instructions to avoid world knowledge inference. Note that the human and LLM annotation protocols were not fully aligned: human annotators followed a structured AIS-based framework restricting inferences to textual evidence only, whereas the LLM received a minimal zero-shot prompt without equivalent framing. This asymmetry likely contributes to the observed disagreements.

To evaluate the fluency of the generated hypotheses, we implemented a structured annotation protocol focused on native-level Farsi linguistic quality. Through manual evaluation of randomly sampled hypotheses, we identified common error patterns in the generated outputs and incorporated these insights into the prompt refinement process. Each hypothesis was labeled as either *Fluent* or *Not fluent*—without intermediate categories or additional explanations.

4.4. LEXICAL ANALYSIS OF SYNTHETIC AND GOLD DATA

Category	Fluent	Non-Fluent
Entailment	39670	16190
Neutral	18110	9308
Contradictions	8520	7499
Total	~ 67k	~33k

Table 4.3: Distribution of labels across Fluent and Non-fluent data

Fluency check prompt

System prompt: You are a helpful assistant to evaluate the fluency of text written in Persian (Farsi).

User prompt: For the sentence below:

- (1) Check if it is grammatically correct, coherent, and sounds natural to a native Persian speaker. You should check for grammar issues, awkward phrasing, broken structure, unclear meaning, hallucinated terms, unnatural repetition, logical errors, etc.
- (2) Think carefully the meaning of the sentence before labeling.
- (3) Label it as either: Fluent or Not fluent. Do not explain.
- (4) Do not fix or rewrite the sentence; only label.

Sentence: {hypothesis}

Table 4.3 presents the distribution of Fluent vs. Non-fluent samples across NLI labels. Approximately 67% of generated hypotheses were classified as fluent, indicating reasonable overall quality. However, fluency varies systematically by label: Entailment hypotheses are most fluent (71%), followed by Neutral (66%), with Contradiction hypotheses showing the lowest fluency (53%). This pattern directly reflects our generation strategy. The mT5-base summarization model produces fluent, faithful summaries when functioning correctly—these naturally correspond to Entailment labels. When the model makes errors (hallucinations, factual mistakes, irrelevant content), outputs are both less fluent and less faithful, yielding Contradiction or Neutral labels. Thus, generation errors serve a dual purpose: they provide negative training examples and signal lower linguistic quality.

Fluency further moderates agreement: for grammatically well-formed premise-hypothesis pairs, agreement rises to 78.9%, whereas degraded or semantically vague pairs reduce agreement to 62.1%. This 16.8 percentage-point gap indicates that annotation noise in the synthetic dataset is concentrated in lower-quality generated hypotheses and that fluency-based filtering could be an effective post-processing step to improve label reliability. We note, however, that the fluency labels were themselves produced by an LLM-as-judge,

so the fluency signal is approximate rather than ground-truth. We retain the dataset as generated in order to evaluate the lower bound of its effect on downstream tasks, while deferring improvements such as prompt engineering and refinement to future work.

Figure 4.9 shows the alignment between human-annotated labels and GPT-4o-mini’s predictions on synthetic data. The Neutral class has the lowest recall and F1-score, indicating the model frequently fails to identify these cases correctly.

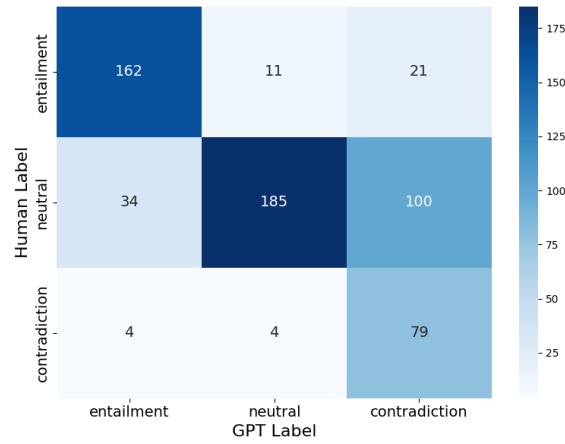


Figure 4.9: Confusion matrix comparing GPT-4o-mini predictions with human-annotated labels on a synthetic dataset of 600 samples.

4.5 Experimental Design

We used mT5 models (mT5-base and mT5-small) trained on the ParsiNLU training dataset as our baseline NLI models following ParsiNLU benchmarks [66]. While we acknowledge more capable models exist (e.g., reasoning-enhanced LLMs), we deliberately selected smaller models without specialized reasoning capabilities as our baseline models to isolate the impact of synthetic data quality from advanced model behaviors^e.

mT5 is the multilingual variant of the T5 ("Text-to-Text Transfer Transformer") [155]. It is a widely used multilingual encoder-decoder model with state-of-the-art performance across many languages, including Farsi^f. These models use short task prefixes to distinguish task intentions. The text-to-text framework enables easier transfer learning evaluation with the same model on a diverse set of tasks. For instance, for MNLI classification, the input might take the form: "mnli hypothesis: I hate pigeons. premise: My feelings towards pigeons are filled with animosity.", and the correspond-

^eExploration of larger instruction-tuned models remains valuable future work, particularly for measuring capability, scaling effects, and synthetic data utility in low-resource settings.

^fmT5 training data includes approximately 52 billion tokens of Farsi text, representing 1.67% of its total training data

ing output is the target label: "Entailment". However, in such text-to-text formulations, the decoder is prone to generating invalid labels outside the predefined label space. We observed instances where the model produced erroneous outputs like "entaildiction", a non-existent word that appeared to be a blend of "Entailment" and "Contradiction". This behavior revealed limitations in sequence generation for label prediction, especially when target labels spanned multiple sub-word tokens. This challenge might arise from the large vocabulary size of the model (250k tokens), which leads to soft output distributions: the model must select among all vocabulary items rather than only among three valid class labels, increasing the likelihood of near-equal probabilities and misclassifications. To mitigate these issues, we use mT5 for classification by adding a classification head on top of the encoder. For sentence-level classification with mT5, the model utilizes the encoder's final hidden states. Each input sequence is expected to end with the special $\langle eos \rangle$ token. The hidden state corresponding to the last $\langle eos \rangle$ token is extracted and treated as the sentence representation. This representation is then passed through a classification head to produce the final logits over the class labels^g. Although this design sacrifices some of the model's full capacity (Table 4.4), it aligns better with the nature of classification task and yields more stable predictions.

Model	Encoder-Decoder	Encoder-Only
mT5-small	300M	172M
mT5-base	540M	390M

Table 4.4: Parameter counts for mT5-small and mT5-base in both encoder-decoder and encoder-only variant with classification head

Hyperparameters: The initial experiment involved fine-tuning the model on the gold-standard dataset. We used a learning rate of $1e-4$ and a batch size of 32. Optimization was performed using the Adam optimizer with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. For regularization, we applied a dropout rate of 0.3 on the classification head and a weight decay of 0.01. We set encoder/decoder dropout to 0, as preliminary experiments showed this drop-out degraded performance. To promote stable training and effective gradient flow, we initialize the linear layers of the classification head using Xavier uniform initialization [42]. This helps prevent vanishing or exploding gradients during early training. In addition, all bias terms are initialized to zero to avoid introducing random shifts in the activations. We train for a maximum of 10 epochs, evaluate checkpoints on steps, and choose the checkpoint with the best ROC-AUC (Receiver Operating Characteristic – Area

^gThis is a standard class of T5 models: https://huggingface.co/docs/transformers/main/en/model_doc/mt5#transformers.MT5ForSequenceClassification

Under the Curve) on the gold validation dataset. In multi-class classification, ROC-AUC (One-vs-Rest) computes a separate ROC curve for each class against the rest, providing a threshold-independent and class-sensitive performance measure. This allows us to assess how well the model discriminates each class independently, making it particularly suitable for selecting the most generalizable checkpoint across diverse training configurations.

Training Regimes and Evaluation: We applied the same fine-tuning procedure across multiple training configurations, including:

- Fine-tuning solely on the synthetic dataset,
- Fine-tuning on a merged set of synthetic and gold-standard data,
- Warming up the model on synthetic data followed by further fine-tuning on gold data^h.

To address class imbalance during training, we employed a weighted loss function that assigns higher importance to underrepresented classes. Furthermore, we created a class-balanced subset of the synthetic dataset and repeated the experiments to isolate the effects of label distribution from data volume. All models were evaluated on the held-out test set derived from the ParsiNLU dataset, ensuring that the reported improvements reflect actual generalization rather than alignment with the annotation model. To control for variance due to stochastic initialization, we ran each experiment with five different random seeds.

GPT-4o-mini zero-shot baseline: To provide the state-of-the-art baseline, we additionally measure the zero-shot performance of GPT-4o-mini on the held-out test set. Without any task-specific fine-tuning, GPT-4o-mini achieves an accuracy of 0.75 and a macro-averaged F1 score of 0.74. Figure 4.10 presents the corresponding confusion matrix. GPT-4o-mini performs strongly on Entailment (83.2% recall) and Neutral (93.5% recall). However, Contradiction is the most error-prone class, with frequent confusion, particularly toward Neutral class.

^hTo complete the training regime comparison, we also evaluated a warmup on gold data followed by synthetic training. This regime consistently underperforms the Synthetic → Gold ordering across both model sizes, confirming that synthetic data is most effective as an initialization signal rather than a fine-tuning target. Based on this observation, Gold → Synthetic experiments were not extended.

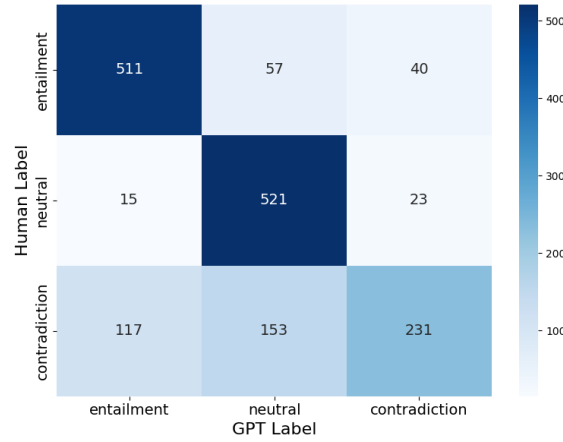


Figure 4.10: Zero-shot classification performance of GPT-4o-mini on the ParsiNLU test set.

4.6 Results

We examined to what extent the synthetic data generated by LLMs can improve the performance on Farsi NLI, by comparing the performance of NLI classifiers trained on the ParsiNLU gold training set with classifiers trained on synthetic data ⁱ.

4.6.1 Synthetic data size effect

We experimented with varying sizes of synthetic data to analyze the cumulative gain of our synthetic data generation. We started from a subset of 5000 synthetic data samples and incrementally added synthetic samples and generated different datasets with varying sizes and balanced label distribution. We did our training and evaluation by using each of these synthetic datasets and reported their performance scores.

Figure 4.11 presents the performance of mT5-small (top row) and mT5-base (bottom row) models across three training strategies and increasing synthetic data sizes (from 5k to around 50k). We found that the cumulative gain from synthetic data is very high when the synthetic dataset size is less than 10k (a steep increase in performance), and after that, the gain is moderate (the performance curve is increasing slowly). If we compare mT5-base and mT5-small, we can see that in mT5-base, a relatively larger model, the performance improvement is plateauing; however, on mT5-small, the performance is still increasing. This pattern can be attributed to the greater learning capacity of the larger model: mT5-base is able to extract useful patterns from the synthetic data more efficiently, reaching diminishing returns sooner. In contrast, mT5-small, with its

ⁱWe were not able to reproduce the reported ParsiNLU benchmark results (0.51 for mT5-small and 0.57 for mT5-base) under our experimental setup on the gold dataset.

more limited capacity, requires a larger volume of synthetic data to capture the same distributional patterns, and thus continues to benefit from additional examples over a wider range of dataset sizes. The lexical saturation observed in Section 4.4 suggests that, beyond a certain scale, additional synthetic examples contribute less novel vocabulary relevant to the gold distribution.

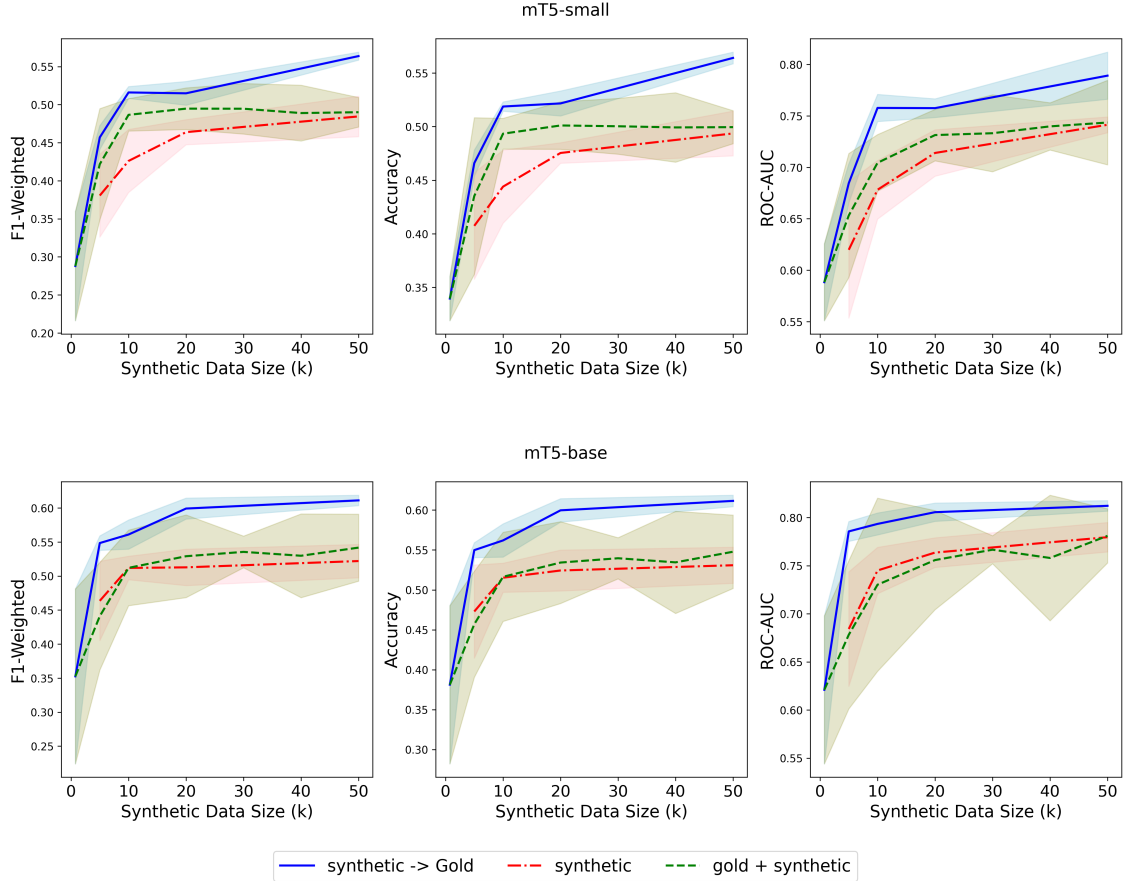


Figure 4.11: Performance of mT5-small and mT5-base models across increasing synthetic data sizes (5k–50k samples) under three training configurations: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded regions indicate 95% confidence intervals over five random seeds. All metrics—weighted F1, accuracy, and ROC-AUC—improve with larger synthetic datasets, with the synthetic→gold transfer achieving the highest and most stable performance across scales.

4.6.2 Training regime effect

From our experimental results in Figure 4.11, we observed that the third training regime, i.e., warming up the model on synthetic data followed by further fine-tuning on gold data (Synthetic→Gold), achieves the overall best performance on all evaluation metrics.

Moreover, one can see that this two-stage strategy (blue) consistently outperforms the other two (red and green) for all dataset sizes. This indicates that our synthetic data is particularly effective as a pre-training signal rather than as a direct replacement for gold data.

For both model sizes, the synthetic-only model lags behind other setups, especially at smaller data sizes, demonstrating limited generalization when synthetic data is used in isolation. However, synthetic-only training outperforms gold-standard training. The merged approach performs competitively in mid-range data sizes (e.g., 10k–30k) but tends to plateau earlier, suggesting that a naive combination of synthetic and gold data may dilute the value of gold supervision. In addition, the Gold $\rightarrow$ Synthetic setting performs similarly to the merged scenario but falls short of the Synthetic $\rightarrow$ Gold curriculum, suggesting that the order of training matters, with synthetic pre-training followed by gold fine-tuning yielding the strongest performance. Notably, the performance gap between training strategies is more pronounced for mT5-small, highlighting that smaller models benefit more from structured training regimes (e.g., warmup followed by fine-tuning). These results underscore the importance of designing training strategies, especially when operating under resource constraints. Finally, we observed that the warm-up training with synthetic data led to more stable optimization dynamics, characterized by smoother gradient norms with lower magnitudes. This stability contributed to faster convergence, with the best ROC-AUC scores often achieved within the early epochs of training. Moreover, the observed gains from the warm-up strategy suggest an implicit form of curriculum learning, where exposure to large-scale, synthetically labeled examples precedes fine-tuning on smaller, gold-standard data.

4.6. RESULTS

Data	Size	ROC	ACC	Weighted-F1
Gold	750	0.603 ± 0.031	0.341 ± 0.004	0.287 ± 0.041
Synthetic(Ours)	5k	0.620 ± 0.046	0.408 ± 0.038	0.380 ± 0.038
	10k	0.678 ± 0.018	0.444 ± 0.024	0.426 ± 0.031
	20k	0.714 ± 0.020	0.464 ± 0.010	0.475 ± 0.009
	50k	0.742 ± 0.008	0.494 ± 0.019	0.485 ± 0.021
	5k+750	0.653 ± 0.060	0.435 ± 0.073	0.422 ± 0.073
Synthetic(Ours) + Gold	10k+750	0.704 ± 0.027	0.493 ± 0.014	0.487 ± 0.021
	20k+750	0.731 ± 0.025	0.501 ± 0.022	0.495 ± 0.027
	30k+750	0.733 ± 0.037	0.500 ± 0.026	0.495 ± 0.033
	40k+750	0.740 ± 0.023	0.499 ± 0.032	0.489 ± 0.037
	50k+750	0.744 ± 0.041	0.500 ± 0.015	0.490 ± 0.019
Synthetic(Ours) → Gold	5k→ 750	0.685 ± 0.019	0.466 ± 0.012	0.457 ± 0.016
	10k→ 750	0.758 ± 0.013	0.519 ± 0.004	0.516 ± 0.008
	20k→ 750	0.758 ± 0.009	0.522 ± 0.012	0.515 ± 0.016
	50k→ 750	0.800 ± 0.051	0.564 ± 0.005	0.564 ± 0.005
Gold → Synthetic(Ours)	750→ 50k	0.741 ± 0.011	0.504 ± 0.017	0.498 ± 0.008

Table 4.5: Performance of **mT5-small** across different dataset sizes (Gold, Synthetic only, Synthetic+Gold, Synthetic→Gold, Gold →Synthetic). Scores are reported as mean $\pm$ 95% CI. Best results per section are highlighted in bold.

4.6. RESULTS

Data	Size	ROC	ACC	Weighted-F1
Gold	750	0.621 ± 0.077	0.381 ± 0.099	0.353 ± 0.129
Synthetic(Ours)	5k	0.684 ± 0.023	0.473 ± 0.046	0.464 ± 0.043
	10k	0.745 ± 0.021	0.515 ± 0.013	0.512 ± 0.013
	20k	0.764 ± 0.012	0.524 ± 0.016	0.513 ± 0.018
	50k	0.780 ± 0.013	0.531 ± 0.014	0.522 ± 0.016
	5k+750	0.678 ± 0.077	0.457 ± 0.066	0.441 ± 0.079
Synthetic(Ours) + Gold	10k+750	0.730 ± 0.090	0.516 ± 0.056	0.512 ± 0.056
	20k+750	0.756 ± 0.052	0.534 ± 0.051	0.529 ± 0.061
	30k+750	0.767 ± 0.015	0.540 ± 0.026	0.536 ± 0.023
	40k+750	0.758 ± 0.065	0.535 ± 0.064	0.530 ± 0.062
	50k+750	0.781 ± 0.028	0.548 ± 0.046	0.542 ± 0.049
Synthetic(Ours) → Gold	5k → 750	0.786 ± 0.010	0.550 ± 0.009	0.549 ± 0.011
	10k → 750	0.794 ± 0.017	0.556 ± 0.020	0.556 ± 0.023
	20k → 750	0.806 ± 0.009	0.600 ± 0.015	0.599 ± 0.015
	50k → 750	0.812 ± 0.005	0.611 ± 0.008	0.612 ± 0.007
Gold → Synthetic(Ours)	750 → 50k	0.782 ± 0.013	0.564 ± 0.014	0.562 ± 0.014

Table 4.6: Performance of **mT5-base** across different dataset sizes (Gold, Synthetic only, Synthetic+Gold, Synthetic→Gold, Gold →Synthetic). Scores are reported as mean $\pm$ 95% CI. Best results per section are highlighted in bold.

While a direct comparison with the ParsiNLU benchmark is not strictly valid due to architectural differences (our encoder-only setup versus their sequence-to-sequence formulation), the relative improvements observed in our experiments are larger (Table 4.7). In the ParsiNLU benchmark, 500k English NLI samples (SNLI) are incorporated by merging them with the gold data. In contrast, training with 50k synthetic samples in our setting yields substantially larger gains, with improvements of +0.14 to +0.16 in the Synthetic+Gold setting and +0.22 to +0.23 in the Synthetic→Gold curriculum, compared to +0.05 to +0.06 when incorporating SNLI. In addition, these gains are achieved with smaller models, suggesting that, in this setting, synthetic data in the target language can be more effective than scaling with large external datasets.

Figures 4.12, 4.13, and 4.14 present the classification results across all three labels: while models trained purely on synthetic data suffer from performance instability—specially on subtle distinctions like Neutral—those trained with a synthetic followed by gold fine-tuning consistently outperform other regimes across all classes. These findings

4.6. RESULTS

Source	Model	Train Data	Acc	Δ
ParsiNLU	mT5-small (~ 300 M)	Gold	0.51	–
		SNLI(500k)+Gold	0.56	+0.05
	mT5-base (~ 540 M)	Gold	0.57	–
		SNLI(500k)+Gold	0.63	+0.06
Ours	mT5-small (~ 170 M)	Gold	0.34	–
		Synthetic(50k)+ Gold	0.50	+0.14
		Synthetic(50k) $\rightarrow$ Gold	0.56	+0.22
	mT5-base (~ 390 M)	Gold	0.38	–
		Synthetic(50k)+ Gold	0.54	+0.16
		Synthetic(50k) $\rightarrow$ Gold	0.61	+0.23

Table 4.7: Comparison with ParsiNLU benchmark. Δ indicates improvement over the gold-only baseline for each model.

provide compelling evidence that strategic integration of synthetic and gold-standard data—specifically through sequential training phases rather than naive data merging—represents the most effective approach.

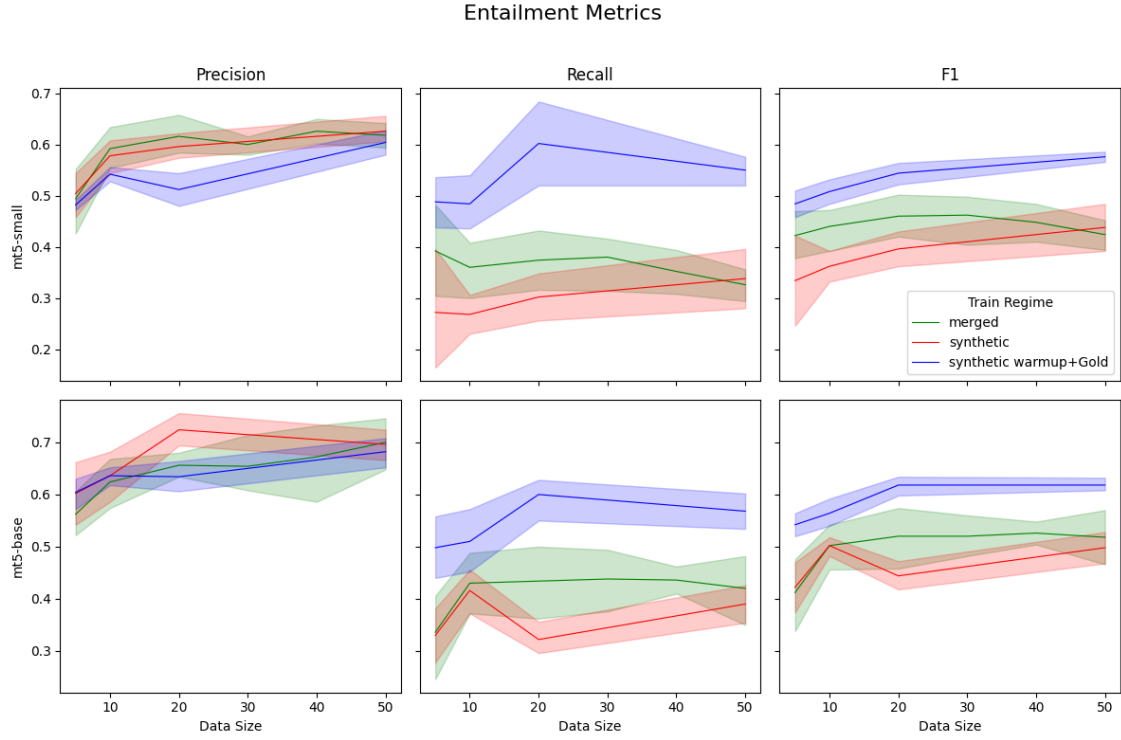


Figure 4.12: Entailment class performance for mT5-small (top row) and mT5-base (bottom row) models across different synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded regions denote 95% confidence intervals over five random seeds. Precision, recall, and F1 consistently improve with larger data, with the Synthetic $\rightarrow$ Gold regime yielding the highest and most stable performance, particularly in recall and overall F1.

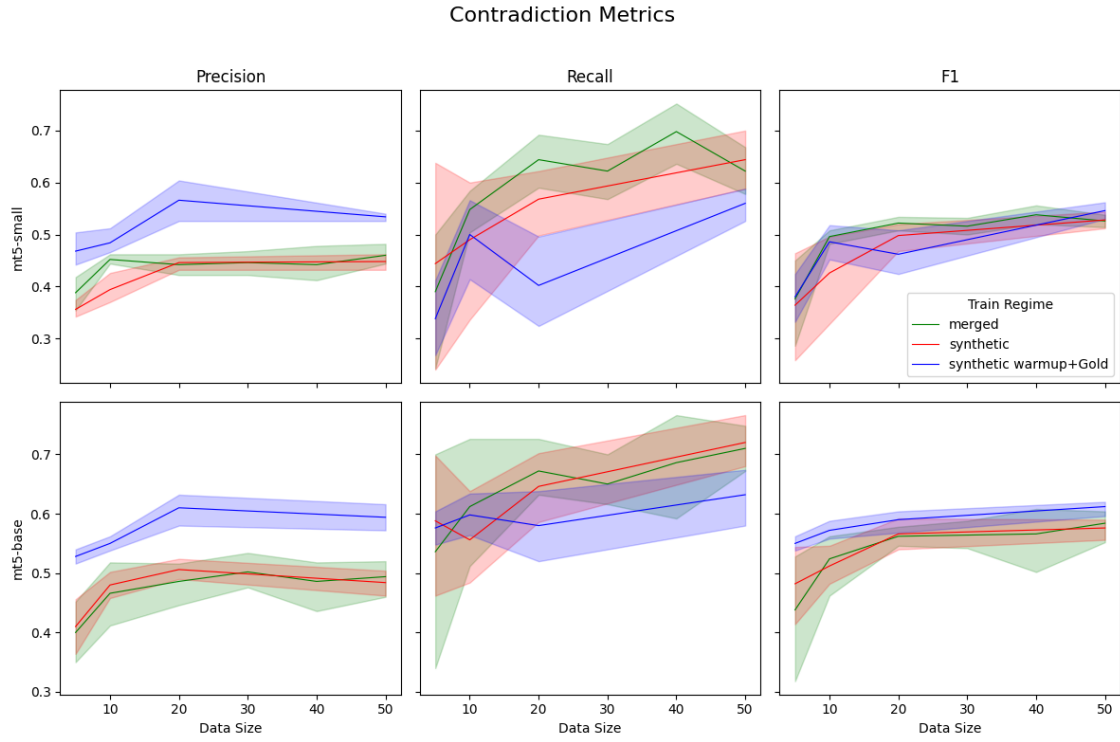


Figure 4.13: Contradiction class performance for mT5-small (top row) and mT5-base (bottom row) models across increasing synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded regions denote 95% confidence intervals over five random seeds. Precision and F1 improve steadily with larger data, while recall shows higher variability across seeds. The Synthetic → Gold regime achieves the most consistent and overall strongest performance, particularly for the mT5-base model.

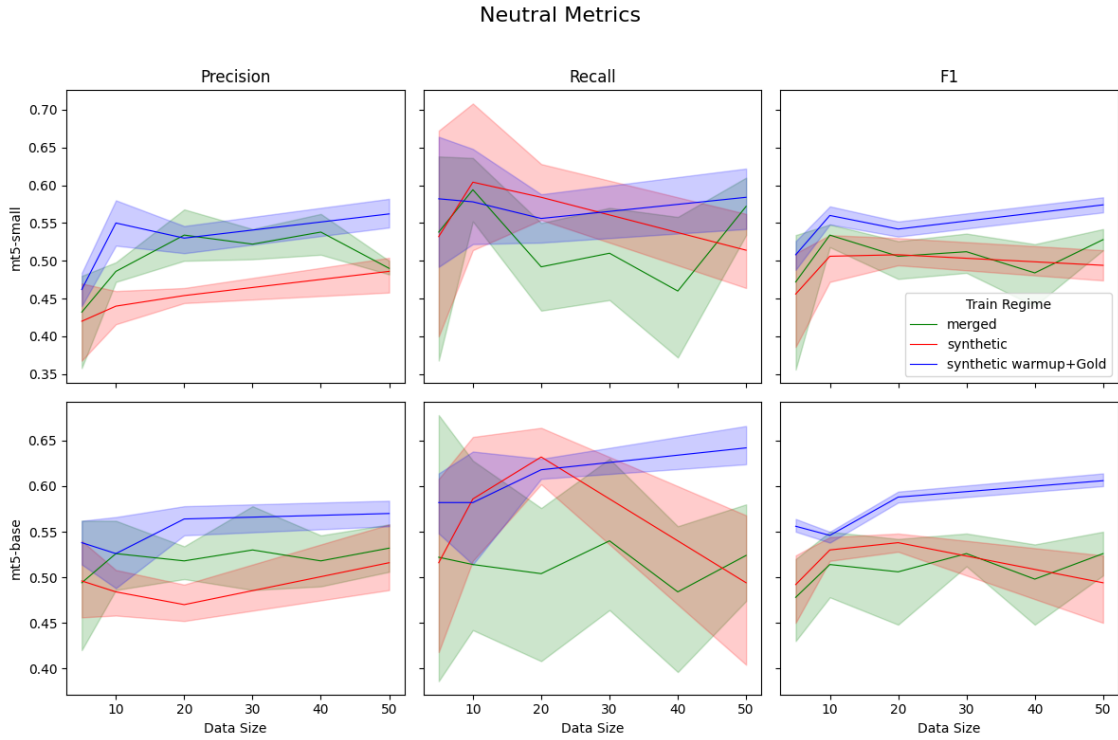


Figure 4.14: Neutral class performance for mT5-small (top row) and mT5-base (bottom row) models across increasing synthetic data sizes (5k–50k) under three training regimes: synthetic-only (red), gold + synthetic merged (green), and synthetic warm-up followed by gold fine-tuning (blue). Shaded areas represent 95% confidence intervals across five random seeds. Precision and F1 steadily improve with data size for both models, while recall shows higher variance. The Synthetic $\rightarrow$ Gold training regime consistently achieves the highest and most stable performance across metrics, indicating better generalization and calibration for the Neutral class.

4.6.3 Effect of Balancing Label Distribution

The synthetic data generated by our pipeline was initially skewed towards Entailment labels, and we balanced this dataset by filtering. Although we initially generated and labeled more than 100,000 samples, filtering reduced this to around 50,000 balanced samples. One natural question here is what the impact of this balancing is. To address this question, we conducted experiments by using the original dataset without balancing. For a fair comparison with the balanced dataset with k samples, we selected the first k samples from the original imbalanced dataset. We trained models using this imbalanced dataset alongside the ParsiNLU train dataset and evaluated them on the ParsiNLU test dataset, with the same training setting as in previous experiments. We plotted the performance scores using this imbalanced dataset, along with the previous performance scores obtained from the balanced dataset, in Figure 4.15.

From the results, we observed that models trained on imbalanced synthetic data consistently underperform their balanced counterparts at smaller scales. However, imbalanced training recovers at larger data sizes, suggesting that class imbalance has a diminishing negative impact as overall volume increases. This has practical implications for data curation strategies in low-resource settings: optimizing for balance may yield greater returns than simply scaling volume.

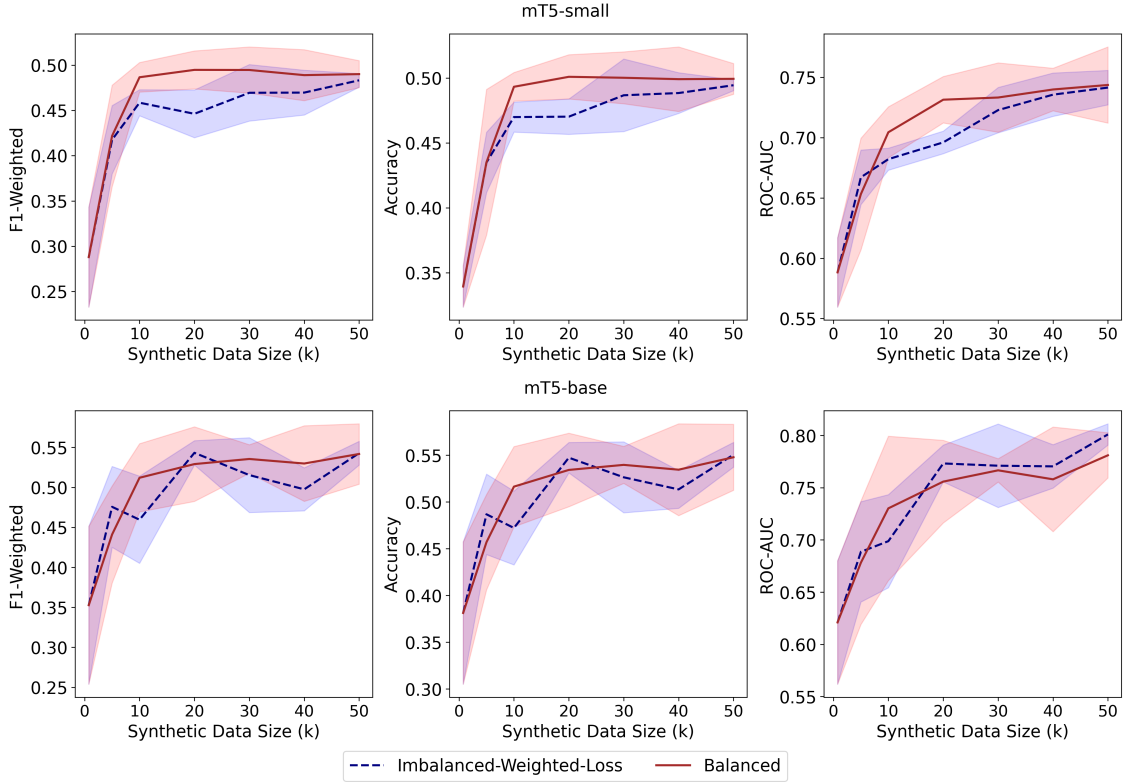


Figure 4.15: Performance of mT5-small (top row) and mT5-base (bottom row) trained on imbalanced and balanced synthetic datasets across increasing data sizes.

4.7 Discussion

The primary objective of our experiments was to assess the general utility of synthetic data in low-resource languages; thus, we did not yet explore all aspects of data curation strategies. Regarding coverage, the input data, sourced exclusively from XLSum’s BBC news corpus, constrains the output to journalistic styles. This stands in contrast to the gold dataset, which includes a more diverse range of linguistic styles. For label quality, we relied solely on zero-shot prompting. However, more robust strategies, such as chain-of-thought prompting, majority voting across different LLMs, or confidence-based filtering, could improve label reliability. Our synthetic data pipeline inherits length constraints

from the summarization model (mT5-base) used for hypothesis generation. To ensure meaningful outputs, we excluded short premises (< 20 tokens) from the source corpus during data creation. This design choice introduces a systematic bias toward longer inputs in the synthetic dataset. Consequently, models trained on this distribution may exhibit degraded performance when evaluated on shorter-premise examples—a form of distributional shift that compromises generalization.

We observed that synthetic data is heavily skewed toward the Entailment class, suggesting that the LLM predominantly generates summaries (hypotheses) faithful to the context. Although smaller generator models, such as mT5-small, produce more non-entailed hypotheses, the annotator model struggles to assign accurate Neutral and Contradiction labels to these outputs, likely because lower-quality hypotheses introduce ambiguity that the annotator cannot reliably resolve. This exposes a fundamental trade-off: increasing class balance through noisier generation degrades label reliability, while prioritizing label quality reinforces class skew. However, high Entailment rates can still be beneficial for certain datasets. For instance, in tasks like summarization, paraphrasing, or faithful sentence generation, the goal is not to distinguish between Entailment, Contradiction, or Neutral, but rather to produce outputs that accurately reflect or preserve the meaning of the original text.

4.8 Summary

In this chapter, we presented a modular, scalable pipeline for generating weakly supervised NLI datasets using LLMs. Our approach demonstrates that even noisy, synthetic data can yield performance gains when integrated appropriately into training workflows. Through our case study in Farsi, we also highlighted practical challenges, including data curation, class imbalance, distributional shifts, and the inherent fuzziness in Entailment judgments. While synthetic data generation offers clear cost and scalability benefits, it also demands careful design choices to ensure quality. In the next chapter, we will investigate the impact of annotator model selection on downstream task performance, examining how it influences the quality and utility of generated datasets.

5

The Role of Label Quality

5.1 Introduction

Building on the demonstrated effectiveness of synthetic data for NLI in the previous chapter, this chapter investigates a key dimension of synthetic data: annotation quality. Chapter 4 established that LLM-generated data can improve model performance but did not address how sensitive downstream performance is to the choice of annotator. As shown in Chapter 3, LLMs can produce various annotations for the same data, raising the question of how this variation affects downstream performance. Traditional human annotation pipelines manage quality through annotator selection, training, guideline design, and adjudication. LLM-based annotation faces analogous challenges: model choice and pre-training data, prompt design, and strategies for aggregating uncertain labels. Understanding how these factors transfer through the synthetic data pipeline is important to developing reliable and efficient data-generation methods. This chapter directly addresses this question through a systematic investigation of the relationship between the reliability of the LLM-annotator and the utility of synthetic data. We structure this investigation around three key aspects. First, we move beyond treating LLMs as a homogeneous class. To isolate the effect of annotator proficiency, we select models with comparable performance on a gold-standard benchmark and evaluate the utility of the synthetic data they produce on a downstream NLI task. This comparative analysis includes both best-prompted generalist models and a specialized NLI model, allowing us to dissect whether raw benchmark performance or inherent task specialization is a stronger predictor of synthetic data quality. Second, we analyze the role of prompting. Previous research, such as [51], has shown that prompt engineering can transform LLMs such as GPT-3.5 into effective annotators that rival human crowdsourced work when provided with sufficient guidance. However, it is not known how these prompt-induced improvements translate into downstream task performance. What remains unclear is how strongly annotation prompt choices translate into downstream model performance once

the data is used for training. To quantify this, we explicitly investigate the effect by using the same model with its best- and worst-performing prompts to generate datasets. The worst-prompt condition is included not as a practical setting but as a lower-bound control to examine the robustness of downstream training to degraded annotation quality. This controlled setup isolates the effect of prompting, enabling us to quantify how annotation quality affects downstream performance independently of the model. Third, we explore the collective-annotation paradigm to enhance label quality, moving beyond single-shot decisions. Inspired by human annotation, which leverages inter-subjective agreement, we evaluate two strategies for aggregating model outputs. (1) a prompt-based ensemble, which utilizes an ensemble across the responses generated by a single LLM to various prompts; (2) a model-based ensemble, which entails aggregating responses from multiple LLMs to a single prompt [143, 135].

The remainder of this chapter is organized as follows. In Section 5.2, we describe our experimental design and provide the method we used. Section 5.3 presents our experimental findings and discusses their implications in Section 5.4. Finally, in Section 5.5, we summarize the findings.

5.2 Method

We systematically investigate how variations in LLM annotator quality influence downstream model performance by comparing synthetic datasets generated under controlled quality conditions. To ensure a fair assessment, all training configurations—including hyperparameters, dataset sizes, and evaluation protocols—are kept constant, as described in Chapter 4. We maintained the same 50,000 premise-hypothesis pairs across all experimental conditions, with only the annotation source varying. This controlled design ensures that any performance differences observed in downstream tasks can be attributed solely to annotation quality rather than to variations in input text. Our analysis considers three complementary dimensions of labeling quality: model choice, prompt design, and ensemble aggregation. Each of these factors is explained in the following sections.

5.2.1 Model Effect: General-Purpose vs Task-Specific Annotators

To assess how the choice of annotator affects downstream performance, we compare three types of models that represent different points in the design space.

- GPT-4o-mini: A cost-effective state-of-the-art proprietary model serving as our baseline.

- Aya-23-35B: A leading open-weight multilingual model with explicit Farsi language coverage.
- DeBERTa-based NLI model: A specialized task-specific model fine-tuned exclusively for natural language inference using the approach described by [72]. The model was trained on multiple NLI datasets, including MNLI, ANLI, and multilingual data. The model achieves an accuracy of 0.74 on the ParsiNLU benchmark, comparable to Aya-23-35B (0.76) and GPT-4o-mini (0.77).

This selection spans proprietary vs. open-weight, general-purpose vs. task-specific, and varying computational requirements. For GPT-4o-mini and Aya-23-35B, we use their respective best-performing prompt configurations identified in Chapter 3 to ensure optimal performance. The DeBERTa model utilized its native classification head without prompt engineering, consistent with its task-specific design. It also serves as a lightweight and computationally efficient alternative to large-scale LLMs. This contrast is reflected in inference cost: annotating the full dataset with Aya-23-35B requires approximately 2 hours on 2 NVIDIA A100 GPUs, whereas DeBERTa completes the same task in roughly 10 minutes on a single RTX 3060 GPU.

5.2.2 Prompts Effect: Within-Model Impact on Downstream Performance

Building on the reliability analysis established in Chapter 3, we investigate how prompt-induced performance variations propagate to downstream task performance by evaluating Aya-23-35B under two contrasting prompt conditions:

- Aya-best: Uses the optimal prompt configuration (`original_elaborate_en`) that achieved 0.76 accuracy.
- Aya-worst: Uses the most problematic prompt (`reverse_concise_fa`) that yielded only 0.61 accuracy.

This controlled comparison enables precise measurement of how prompt quality—independent of model, architecture, and capacity—affects the utility of synthetic training data. By holding the model constant while varying only the prompt formulation, we isolate the specific impact of prompt design choices on downstream learning outcomes. The substantial performance gap between these conditions (15 percentage points in accuracy) provides a clear test case for whether prompt-induced differences in annotation quality translate into measurable downstream effects.

5.2.3 Ensemble Effect: Quality Optimization Strategies

Based on the findings in Chapter 3, we observe variability at both the model and prompt levels, with differences in agreement with human annotations and in class-wise sensitiv-

ity and consistency. Chapter 3 showed that different models excelled on different label classes and that prompt phrasing materially shifted agreement with human annotators—suggesting that no single model-prompt configuration is uniformly optimal. We therefore examine whether aggregation strategies could mitigate individual annotator limitations and improve overall label quality. We implement two complementary ensemble approaches:

- Prompts ensemble(Aya-23-35B): Majority voting across eight prompt variants, leveraging the model’s diverse behavioral profiles under different prompting conditions. On the ParsiNLU dataset, this approach achieves an accuracy of 0.76, comparable to the best single-prompt configuration. For further analysis, we extend this setup by applying prompt-variant aggregation to all models. This allows us to examine whether the relationship observed for Aya-23-35B—where ensemble performance matches the best individual prompt—holds consistently across different models.
- Open-model ensemble: Majority voting across three open-weight models, capturing model family diversity and complementary strengths. On ParsiNLU, the Models-ensemble achieves the highest overall accuracy (0.79), exceeding the performance of the strongest individual model, which attains an accuracy of 0.77.

By comparing ensemble performance with single-model baselines, we can quantify the potential gains from multi-annotator aggregation and identify scenarios where downstream performance improvements justify the additional computational overhead of ensemble methods.

5.3 Results

5.3.1 Annotator Bias: Label Distribution Shifts Across Models and Prompts

Starting from the balanced 50k subset labeled by GPT-4o-mini in Chapter 4, we re-annotated the same premise–hypothesis pairs with different additional configurations to isolate the effect of annotator choice on label distribution. We compare four single-annotator configurations—GPT-4o-mini, Aya-23-35B under its best- and worst-performing prompts from Chapter 3, and DeBERTa-NLI—alongside two ensemble strategies that aggregate across prompts and across models, respectively. Table 5.1 reports the resulting class counts on the 50k-sample synthetic dataset.

While GPT-4o-mini’s original annotations showed a natural skew toward Entailment, we used the balanced subset from Chapter 4 (16k examples per class) to maintain experimental consistency and enable direct comparison across chapters. However, Aya-23-35B under its optimal prompt configuration exhibited Entailment skew (25k Entailment vs.

5.3. RESULTS

Model	Entailment	Neutral	Contradiction
GPT-4o-mini (best-prompt)	16k	16k	16k
Aya-23-35B (best-prompt)	25k	16k	8k
Aya-23-35B (worst prompt)	39k	6k	5k
DeBERTa-NLI	20k	18k	11k
Prompts-ensemble	28k	12k	9k
Models-ensemble	16k	19k	14k

Table 5.1: Label distribution across synthetic datasets generated by different annotator configurations. Notable variations in class balance reflect model-specific annotation biases.

16k Neutral and 8k Contradiction). This imbalance becomes particularly pronounced under suboptimal prompting, with the worst-prompt condition producing severe Entailment dominance (39k examples) and drastic underrepresentation of Contradiction (5k) and Neutral (6k) classes. The DeBERTa model demonstrated a more moderate skew toward Entailment (20k) while maintaining better representation of Neutral (18k) and Contradiction (11k) classes, achieving a moderate agreement of 0.55 with GPT-4o-mini.

Figure 5.1 presents a pairwise comparison of label alignment between GPT-4o-mini and other annotators: Aya-worst, Aya-best, and DeBERTa-NLI across the three NLI classes. The Aya-worst model exhibits a strong bias toward Entailment, producing highly skewed predictions and limited coverage of the Neutral and Contradiction classes. This confirms that suboptimal prompting can reduce the diversity and informativeness of synthetic labels. By contrast, Aya-best shows substantially improved alignment with GPT-4o-mini across all three classes, particularly in the Neutral category, indicating that better prompts yield more balanced, semantically grounded annotations. The DeBERTa-NLI model, while not instruction-tuned and substantially smaller, exhibits decent agreement with GPT-4o-mini.

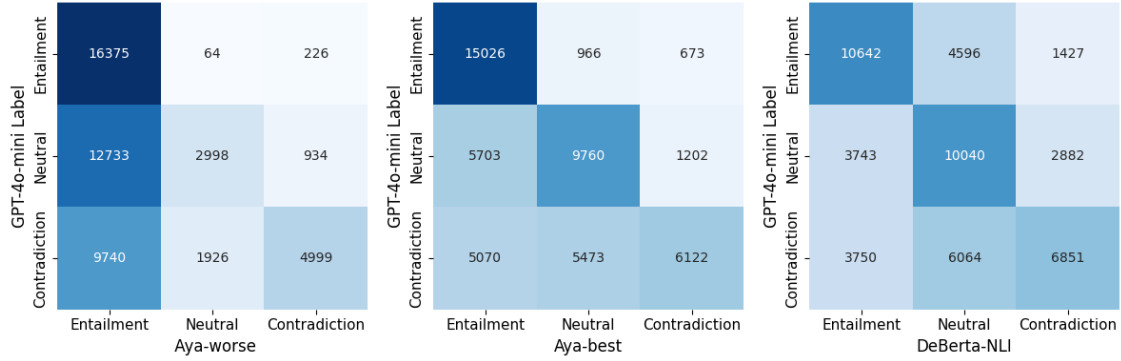


Figure 5.1: Comparison of pairwise confusion matrices between GPT-4o-mini and other annotators (Aya-worst, Aya-best, and DeBERTa-NLI) across NLI classes. GPT-4o-mini and Aya-best show the strongest alignment, especially for *Entailment* and *Neutral*. Aya-worst and DeBERTa-NLI display larger off-diagonal errors, indicating greater divergence in annotation behavior and decision boundaries.

Aggregating eight prompt variants of the same model (Aya-23-35B) yields a distribution of 28k Entailment, 12k Neutral, and 9k Contradiction—slightly more skewed toward Entailment than Aya’s single best-prompt configuration (25k / 16k / 8k). This is consistent with the prompt-sensitivity analysis in Chapter 3: Aya-23-35B exhibits lower prompt sensitivity on Entailment and higher class-internal consistency for Entailment than for Neutral, meaning its predictions on Entailment examples are stable across prompt variants while its Neutral predictions are not. Majority voting across prompt variants therefore reinforces a systematic, model-specific tendency to over-predict Entailment, and the high confusion between Neutral and Entailment observed in Figure 3.8 amplifies rather than corrects this imbalance. The prompts-ensemble result is consistent with the observed mean pairwise agreement (≈ 0.75) and Cohen’s κ values ranging from 0.65 to 0.85 across the eight prompt variants (Figure 3.10). Given this relatively high agreement, majority voting across prompts provides limited additional information and is therefore unlikely to resolve systematic disagreements between prompt formulations. As a result, the prompts-ensemble shows relatively low agreement with GPT-4o-mini (0.44), indicating that aggregation does not improve alignment with the reference annotator.

The cross-model ensemble—combining Aya-23-35B, Qwen3-30B, and Llama-3.3-70B—produces the most balanced distribution observed in this study (16k / 19k / 14k), close to GPT-4o-mini’s reference distribution. This behavior is consistent with the observations in Chapter 3: Qwen3-30B and Llama-3.3-70B exhibit similar per-class behavior and achieve stronger performance on the *Contradiction* class compared to Aya-23-35B (Figure 3.2). In addition, under their best prompt configurations, Qwen3-30B and Llama-3.3-70B show substantially higher agreement with GPT-4o-mini (≈ 0.80) than Aya-23-35B (≈ 0.60) (Figure 3.10). Their complementary strengths help counteract Aya’s tendency

toward Entailment, resulting in a more balanced overall label distribution under majority voting. Consistent with this, the model ensemble achieves agreement of 0.68 with GPT-4o-mini on the synthetic dataset, indicating improved alignment relative to Aya alone. Furthermore, inter-annotator agreement among the three open models, measured by Fleiss’ κ , decreases from 0.69 on the gold-standard data (Table 3.9) to 0.50 on the synthetic data, indicating that LLM-generated examples introduce additional ambiguity and reduce labeling consistency.

The asymmetry between the two ensemble strategies is also visible at the per-example level (Figure 5.2). The model-ensemble achieves substantially higher per-example agreement (74% diagonal mass) and produces a roughly symmetric off-diagonal pattern, indicating that disagreements are distributed across classes rather than concentrated in one direction. The prompts-ensemble, by contrast, exhibits a strongly asymmetric error pattern, with a large proportion of *Neutral* and *Contradiction* examples misclassified as *Entailment*.

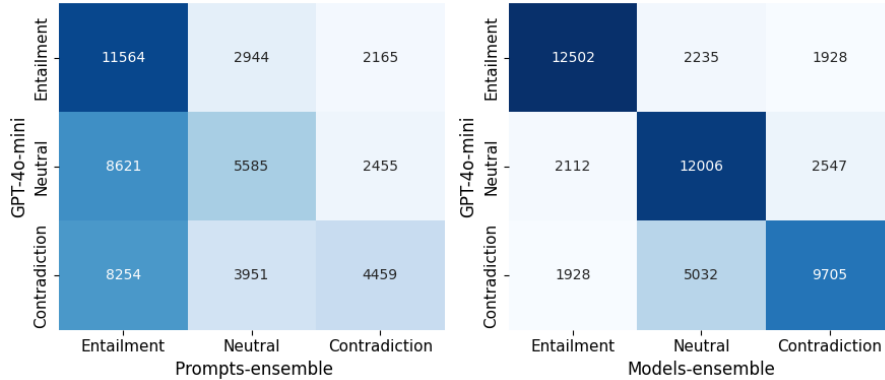


Figure 5.2: Comparison of pairwise confusion matrices between GPT-4o-mini and other annotators (Prompts-ensemble and Models-ensemble) across NLI classes. GPT-4o-mini and Models-ensemble exhibit the strongest alignment, especially for *Entailment* and *Neutral*. Prompts-ensemble is highly skewed toward Entailment.

These observations have several implications for synthetic dataset construction: (i) label distributions are highly sensitive to annotator choice and prompt formulation; (ii) prompt-induced biases can exacerbate class imbalance, particularly for minority classes such as Contradiction; and (iii) disagreement across annotators suggests that relying on a single model may introduce systematic labeling bias, motivating the use of multi-model aggregation.

5.3.2 Prompt-Variant Insights Across Models

We assess ensemble methods across prompt variations on ParsiNLU, analyzing how aggregation strategies mitigate individual prompt limitations and improve overall annotation stability. Table 5.2 presents the accuracy of each model under eight distinct prompt configurations, alongside the ensemble performance. For cases without a clear majority, we used random tie-breaking to ensure unbiased aggregation. The ensemble approach consistently delivers competitive results, often matching or exceeding the best individual prompt performance for each model.

Model	reverse_elaborate_en	reverse_concise_en	reverse_elaborate_fa	reverse_concise_fa	original_elaborate_en	original_concise_en	original_elaborate_fa	original_concise_fa	Majority
Aya-23-35B	0.64	0.68	0.61	0.64	<u>0.76</u>	0.75	0.70	0.71	0.76
Qwen3-30B	0.70	0.69	0.68	0.65	<u>0.77</u>	0.76	0.74	0.75	0.77
Llama3-70B	0.71	0.70	0.69	0.71	0.77	0.77	0.74	<u>0.78</u>	0.79
GPT-4o-mini	0.73	0.70	0.72	0.68	<u>0.77</u>	0.77	0.76	0.75	0.80
GPT-4o	0.76	0.74	0.75	0.75	0.76	<u>0.79</u>	0.76	0.76	0.82
Claude-Opus-4	0.76	0.74	0.76	0.76	0.79	0.81	0.82	<u>0.83</u>	0.83

Table 5.2: Accuracy of different models under prompt variations with Prompts-ensemble on the ParsiNLU (train) dataset. Results are reported for reverse/original, elaborate/-concise, and English/Persian prompts. The last column shows ensemble accuracy using Models-ensemble with a random tie-break. The number of tied cases is 63, 60, 50, 51, 43, and 32, respectively. Best individual prompt scores are underlined, and the overall best scores are highlighted.

Figure 5.3 illustrates the relationship between individual prompt agreement with majority-vote labels and accuracy on the gold dataset across all eight prompt variants for each model. Across all models, higher agreement with the ensemble is associated with the prompt with higher gold accuracy. This relationship is most pronounced for Claude-Opus-4 ($r=0.96$) and remains substantial for Aya-23-35B ($r=0.87$) and Llama3-70B ($r=0.80$), indicating that majority-vote agreement serves as an excellent proxy for accuracy in these architectures. GPT-4o-mini ($r=0.55$) and GPT-4o ($r=0.51$) exhibit moderate correlations, while Qwen3-30B demonstrates the weakest relationship ($r=0.46$),

reflecting greater performance variability across prompt conditions.

A key implication of these findings is that for most models, the prompt variant with the highest agreement with the ensemble majority vote is also the most accurate. This provides a practical method for selecting optimal prompts when gold labels are unavailable.

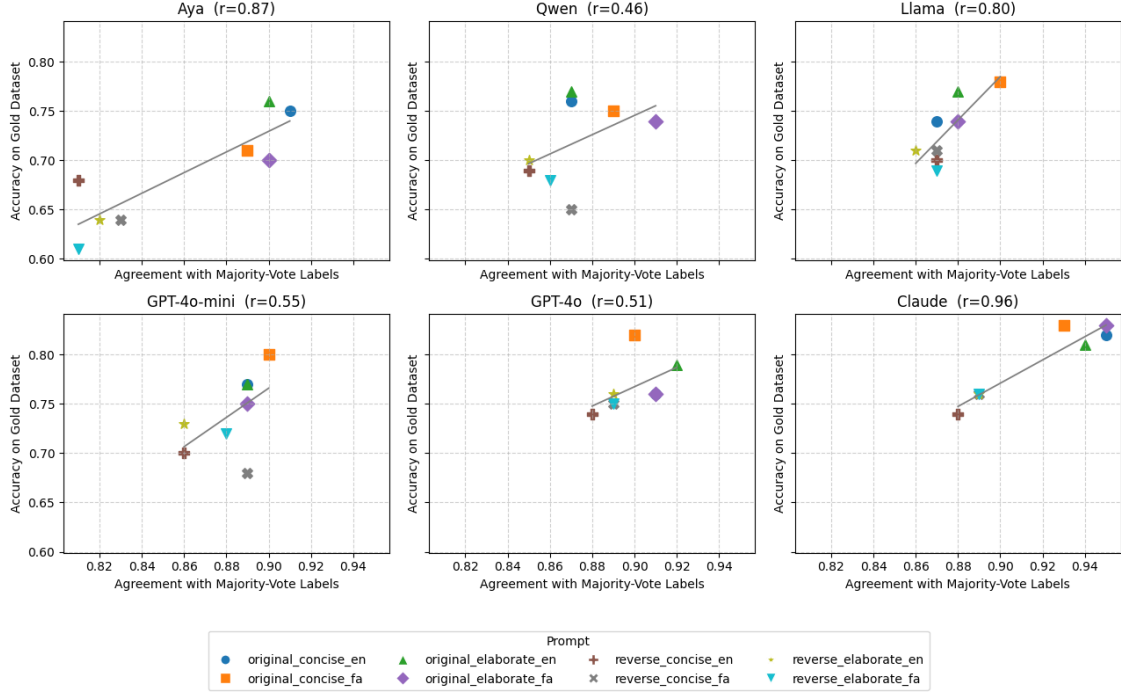


Figure 5.3: Correlation between prompt-level agreement with majority-vote labels and gold-standard accuracy across eight prompt variants for each model. Strong positive correlations for Claude-Opus-4 ($r=0.96$), Aya-23-35B ($r=0.87$), and Llama3-70B ($r=0.80$) indicate that ensemble agreement serves as a reliable proxy for annotation quality. In contrast, weaker correlations for GPT variants ($r=0.51$ - 0.55) and Qwen3-30B ($r=0.46$) reflect greater prompt-induced performance variability. Least-squares regression lines and Pearson correlation coefficients demonstrate the strength of the relationship for each model family.

5.3.3 Downstream Impact of Annotation Choices

We trained student models (mT5-small and mT5-base) on the synthetic data generated using Aya-23-35B under **best** and **worst** prompts and the DeBERTa models, as well as ensemble-based annotations, including the **prompts-ensemble** (Aya-23-35B) and the **models-ensemble** with the same settings as Chapter 4. To mitigate the effects of class imbalance across annotators, we used a weighted loss function during training. To establish a lower bound on performance, we included a random-labeling condition in which

labels were assigned according to the class distribution observed in the gold-standard dataset. This baseline helps contextualize the quality spectrum from completely random annotations to optimized ensemble strategies.

Figures 5.4, 5.5, and 5.6 show the accuracies and weighted F1 scores for each model under three training regimes. The first observation is that prompt-induced annotation noise measurably propagates into downstream performance, and the magnitude of the effect depends on both model scale and training regime. For mT5-small in the synthetic-only setting, the gap between Aya-best and Aya-worst is 8 points; this gap shrinks to 6 points when gold data is added and to 4 points when gold fine-tuning follows synthetic training. The same pattern is more pronounced in mT5-base, where the gaps are 13, 12, and 4 points, respectively. The larger model is more sensitive to upstream annotation quality—a finding consistent with prior work on noise-capacity interactions in language models—but in both cases gold fine-tuning attenuates rather than eliminates the cost of poor annotation. The second observation is that this gap is not driven solely by the model architecture. DeBERTa-NLI, despite being orders of magnitude smaller than the open-weight LLMs and not relying on prompt engineering, produces synthetic data that performs within 2 percentage points of GPT-4o-mini and Aya-best across all training regimes. This suggests that for NLI annotation, task-specific training of a small model can substitute for scale and prompt optimization in a frontier LLM.

The two ensembling strategies yield comparable downstream performance in the synthetic-only and synthetic+gold regimes, with differences of approximately one percentage point across both mT5 scales. Under synthetic $\rightarrow$ gold fine-tuning, however, mT5-small treats them identically (0.544 prompts-ensemble vs. 0.550 model-ensemble accuracy), while mT5-base shows a clear preference for the prompts-ensemble (0.613 vs. 0.591 for model-ensemble). Tables 5.3 and 5.4 provide detailed results.

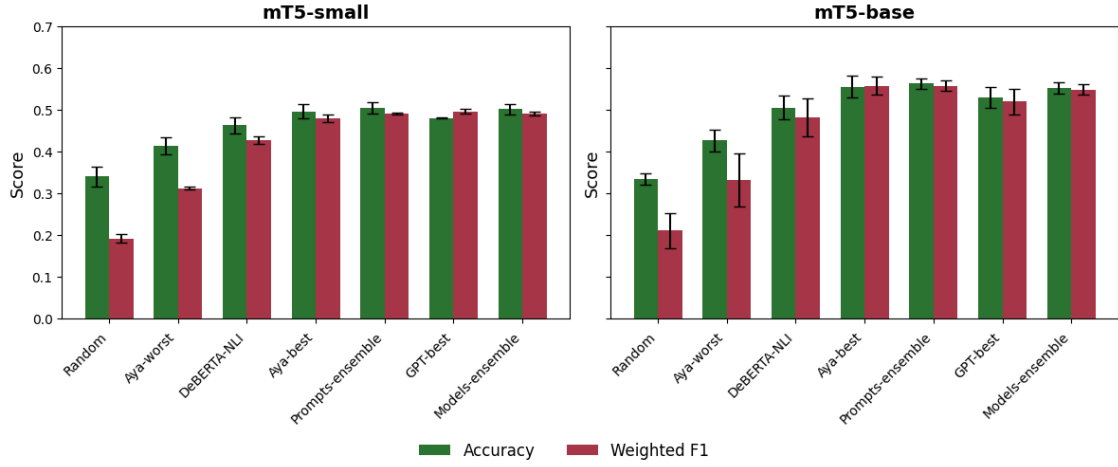


Figure 5.4: Performance of mT5-small and mT5-base under *Synthetic* training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. For both model sizes, random or low-quality annotations (Aya-worst) yield substantially lower scores, confirming the importance of annotation quality.

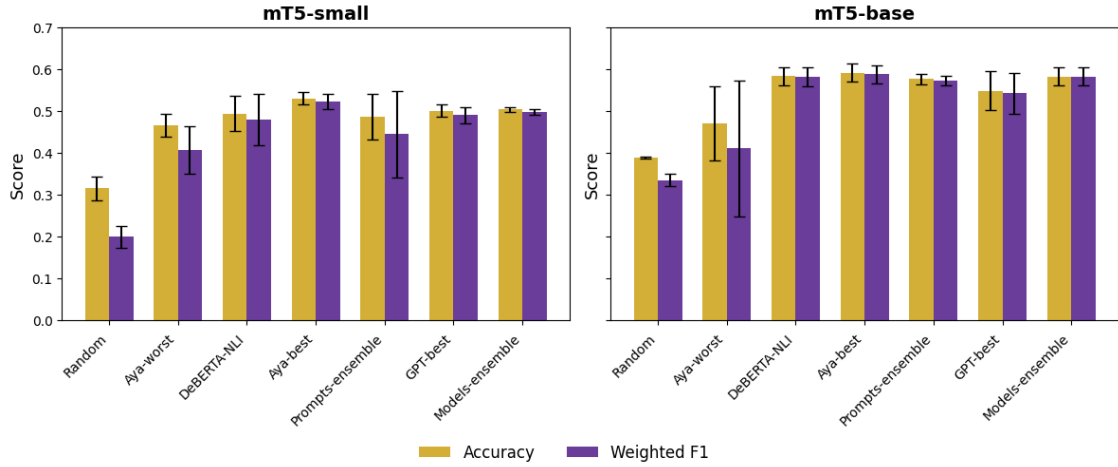


Figure 5.5: Performance of mT5-small and mT5-base under *Synthetic(Ours)+Gold* training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. In the mT5-small model, the Aya-best annotator consistently outperforms all other conditions. In contrast, for mT5-base, the Models-ensemble and DeBERTa-NLI annotators achieve performance comparable to Aya-best.

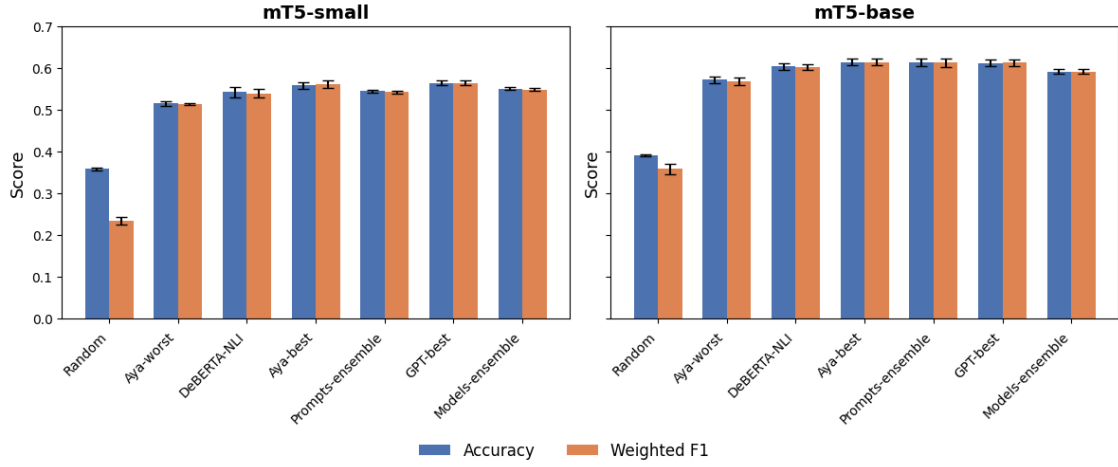


Figure 5.6: Performance of mT5-small and mT5-base under *Synthetic(Ours)* $\rightarrow$ *Gold* training regimes and annotator quality conditions. Bar plots show Accuracy and Weighted F1 (mean $\pm$ 95% CI) for each model. Under synthetic $\rightarrow$ gold fine-tuning, mT5-small treats prompts-ensemble and model-ensemble annotators near-identically, while mT5-base shows a clear preference for the prompts-ensemble over the model-ensemble.

5.3. RESULTS

Data	ROC	ACC	Weighted-F1
Gold	0.603 ± 0.031	0.341 ± 0.004	0.287 ± 0.041
<i>Synthetic(Ours)</i>			
Random	0.564 ± 0.018	0.340 ± 0.024	0.191 ± 0.036
Aya-worst	0.693 ± 0.011	0.413 ± 0.021	0.312 ± 0.052
DeBERTa-NLI	0.661 ± 0.043	0.462 ± 0.020	0.427 ± 0.032
Aya-best	0.720 ± 0.022	0.496 ± 0.016	0.480 ± 0.002
GPT-best	0.744 ± 0.006	0.480 ± 0.002	0.496 ± 0.016
Prompts-ensemble	0.734 ± 0.011	0.504 ± 0.013	0.490 ± 0.022
Models-ensemble	0.714 ± 0.008	0.501 ± 0.012	0.490 ± 0.011
<i>Synthetic(Ours)+Gold</i>			
Random	0.522 ± 0.011	0.315 ± 0.028	0.199 ± 0.026
Aya-worst	0.718 ± 0.029	0.465 ± 0.027	0.406 ± 0.057
DeBERTa-NLI	0.719 ± 0.028	0.493 ± 0.042	0.479 ± 0.061
Aya-best	0.754 ± 0.019	0.530 ± 0.014	0.522 ± 0.019
GPT-best	0.744 ± 0.041	0.500 ± 0.015	0.490 ± 0.019
Prompts-ensemble	0.711 ± 0.058	0.486 ± 0.054	0.444 ± 0.103
Models-ensemble	0.731 ± 0.034	0.503 ± 0.006	0.498 ± 0.007
<i>Synthetic (Ours) $\rightarrow$ Gold</i>			
Random	0.545 ± 0.041	0.358 ± 0.004	0.234 ± 0.010
Aya-worst	0.756 ± 0.007	0.515 ± 0.006	0.513 ± 0.003
DeBERTa-NLI	0.747 ± 0.006	0.542 ± 0.012	0.539 ± 0.010
Aya-best	0.776 ± 0.051	0.558 ± 0.008	0.560 ± 0.009
GPT-best	0.800 ± 0.051	0.564 ± 0.005	0.564 ± 0.005
Prompts-ensemble	0.760 ± 0.006	0.544 ± 0.003	0.542 ± 0.003
Models-ensemble	0.751 ± 0.006	0.550 ± 0.004	0.548 ± 0.004

Table 5.3: mT5-small performance across annotator quality spectrum and training regimes using 50k synthetic data. Results demonstrate the impact of annotation quality from random labels to ensemble methods across three training scenarios: synthetic-only, combined synthetic+gold, and synthetic pre-training followed by gold fine-tuning. Performance metrics (ROC-AUC, Accuracy, Weighted F1) are reported as mean $\pm$ 95% confidence intervals, revealing systematic quality gradients from weakest to strongest annotators.

5.3. RESULTS

Data	ROC	ACC	Weighted-F1
Gold	0.621 ± 0.077	0.381 ± 0.099	0.353 ± 0.129
<i>Synthetic(Ours)</i>			
Random	0.568 ± 0.012	0.334 ± 0.014	0.210 ± 0.043
Aya-worst	0.732 ± 0.027	0.426 ± 0.026	0.331 ± 0.063
DeBERTa-NLI	0.738 ± 0.044	0.505 ± 0.029	0.481 ± 0.045
Aya-best	0.788 ± 0.033	0.555 ± 0.027	0.557 ± 0.022
GPT-best	0.777 ± 0.020	0.529 ± 0.026	0.519 ± 0.030
Prompts-ensemble	0.785 ± 0.012	0.562 ± 0.013	0.557 ± 0.013
Models-ensemble	0.774 ± 0.006	0.552 ± 0.013	0.548 ± 0.013
<i>Synthetic(Ours)+Gold</i>			
Random	0.573 ± 0.024	0.388 ± 0.003	0.334 ± 0.015
Aya-worst	0.748 ± 0.039	0.470 ± 0.089	0.410 ± 0.162
DeBERTa-NLI	0.794 ± 0.018	0.583 ± 0.022	0.582 ± 0.023
Aya-best	0.804 ± 0.011	0.591 ± 0.021	0.587 ± 0.022
GPT-best	0.761 ± 0.028	0.548 ± 0.046	0.542 ± 0.049
Prompts-ensemble	0.781 ± 0.012	0.576 ± 0.012	0.572 ± 0.011
Models-ensemble	0.809 ± 0.009	0.582 ± 0.021	0.582 ± 0.022
<i>Synthetic (Ours) $\rightarrow$ Gold</i>			
Random	0.595 ± 0.011	0.390 ± 0.002	0.358 ± 0.012
Aya-worst	0.798 ± 0.009	0.571 ± 0.007	0.568 ± 0.009
DeBERTa-NLI	0.812 ± 0.003	0.603 ± 0.008	0.602 ± 0.007
Aya-best	0.835 ± 0.006	0.614 ± 0.007	0.614 ± 0.007
GPT-best	0.812 ± 0.005	0.611 ± 0.008	0.612 ± 0.007
Prompts-ensemble	0.836 ± 0.003	0.613 ± 0.009	0.612 ± 0.010
Models-ensemble	0.808 ± 0.007	0.591 ± 0.005	0.591 ± 0.005

Table 5.4: mT5-base performance across annotator quality spectrum and training regimes using 50k synthetic data. Results demonstrate the impact of annotation quality, from random labels to ensemble methods, across three training scenarios: synthetic-only, synthetic+gold, and synthetic pre-training followed by gold fine-tuning. Performance metrics (ROC-AUC, Accuracy, Weighted F1) are shown as mean $\pm$ 95% confidence intervals, revealing systematic quality gradients from weakest to strongest annotators.

Per-Class Results

To better understand how annotation quality affects model behavior across different classes, we examine the per-class results for Entailment, Neutral, and Contradiction.

Figures 5.7 and 5.8 show the detailed per-class Precision, Recall, and F1 scores for the mT5-small and mT5-base models in the sequential training regime.

Entailment Both models achieve the best Entailment performance when trained on high-quality annotations. For mT5-small, the effects of annotation quality are evident in the spread between systems. Models trained on Aya-best and GPT-4o-mini achieve precision of 0.58–0.60 and balanced recall, yielding F1 scores near 0.58. In contrast, Aya-worst yields lower precision and greater confidence intervals. DeBERTa-NLI and Prompts-ensemble achieve intermediate but stable results, suggesting that model-based or ensemble-derived labels provide reasonably consistent supervision, albeit slightly less precise.

For mT5-base, the same pattern persists but with clearer separation among annotators. Models trained on Aya-best and GPT-4o-mini achieve higher precision, confirming that annotation quality directly strengthens Entailment discrimination. Prompts-ensemble matches the stronger systems on average but displays wider uncertainty, consistent with variability in its aggregated labels.

Neutral Unlike Entailment, where gains are primarily seen in precision, the Neutral class exposes differences in recall and overall balance across annotators. High-quality labels from Aya-best, GPT-4o-mini, and DeBERTa-NLI consistently yield smoother precision–recall trade-offs, while noisier sources such as Aya-worst introduce confusion between Neutral and Entailment examples. For mT5-small, models trained on Aya-best and GPT-4o-mini achieve balanced precision and recall, indicating that these annotators provide consistent boundary examples in which Neutrality is properly preserved. Prompts-ensemble and DeBERTa-NLI offer intermediate results with smaller confidence intervals, indicating reliable but slightly less refined representations of Neutrality.

For mT5-base, the same trend is observed, but the distinction among annotators becomes more pronounced. Aya-best achieves the highest recall and a strong F1 score, indicating that its synthetic labels provide clearer signals for Neutral inference boundaries. GPT-4o-mini and Prompts-ensemble follow closely, showing stable, consistent results with similar recall and precision.

Contradiction Across both models, Contradiction appears to be the most robust class to annotation variability. In mT5-small, Aya-worst performs on par with others, suggesting that smaller models are less sensitive to annotation noise. However, in mT5-base, the difference becomes more pronounced, and Aya-worst shows greater variance. This pattern indicates that larger models, being more discriminative, also amplify the effects of poor annotation quality, revealing limitations that smaller models tend to smooth over.

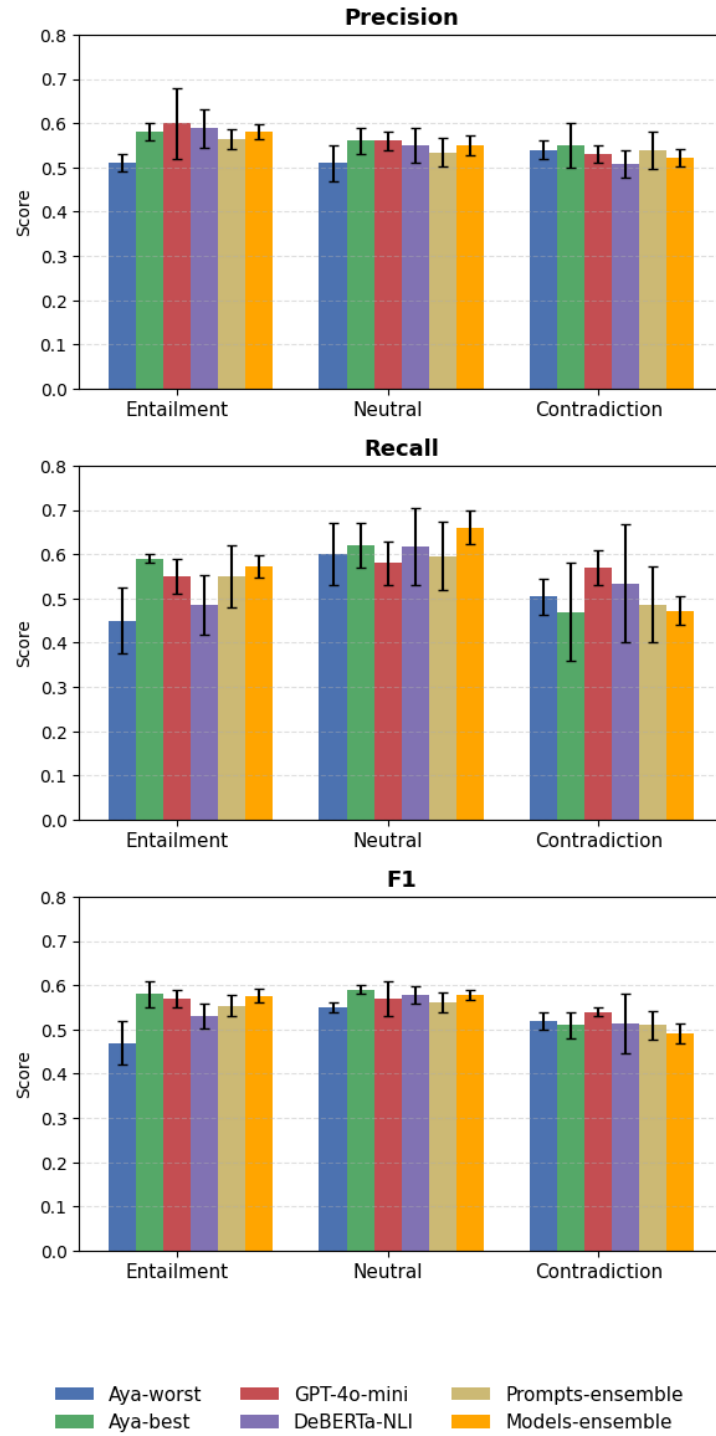


Figure 5.7: Per-Class Performance of mT5-small (Synthetic $\rightarrow$ Gold) across annotator quality conditions.

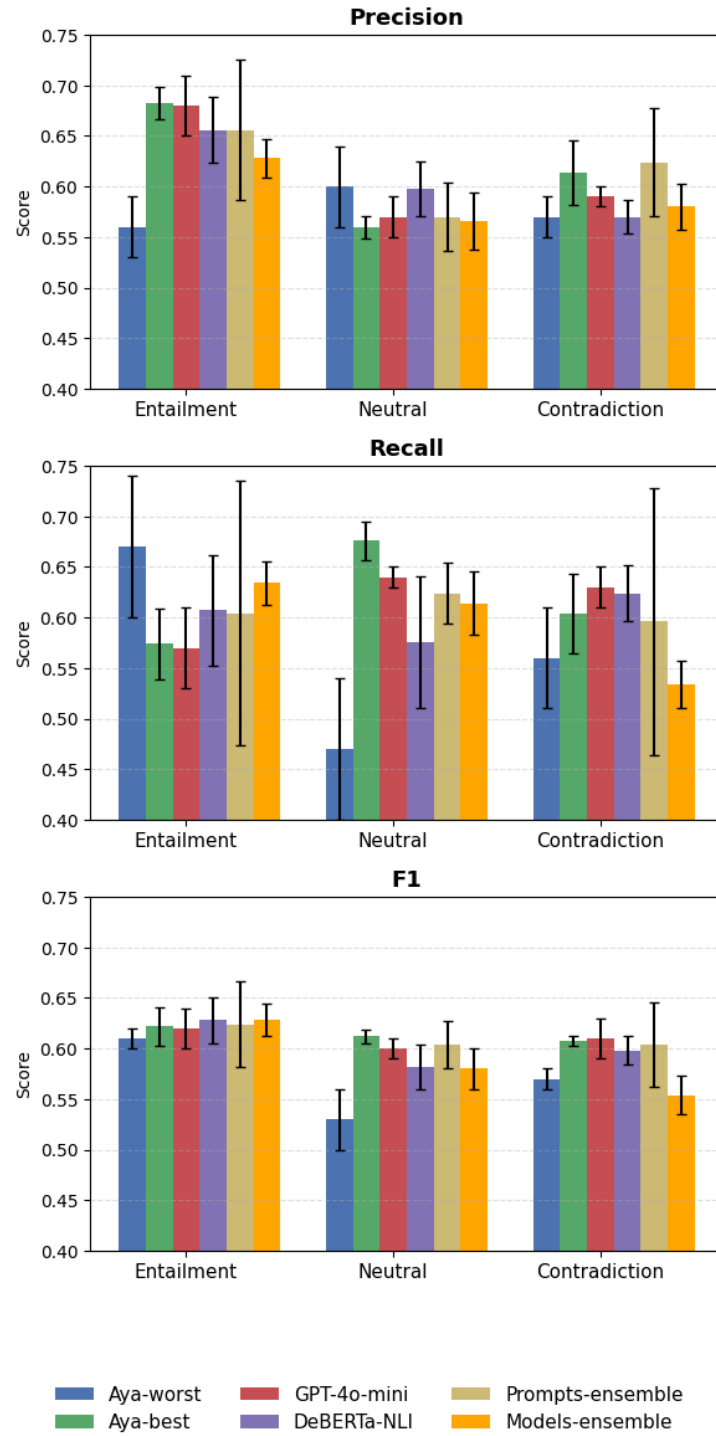


Figure 5.8: Per-Class Performance of mT5-base (Synthetic $\rightarrow$ Gold) across annotator quality conditions.

Tables 5.5 and 5.6 provide the detailed reports on classification.

5.3. RESULTS

Metric	Aya-worst	Aya-best	GPT-4o-mini	DeBERTa-NLI	Prompts ensemble	Models ensemble
<i>Entailment</i>						
Precision	0.51 ± 0.02	0.58 ± 0.02	0.60 ± 0.08	0.588 ± 0.044	0.564 ± 0.022	0.580 ± 0.017
Recall	0.45 ± 0.074	0.59 ± 0.01	0.55 ± 0.04	0.486 ± 0.067	0.55 ± 0.070	0.572 ± 0.025
F1	0.47 ± 0.049	0.58 ± 0.03	0.57 ± 0.02	0.530 ± 0.028	0.554 ± 0.025	0.576 ± 0.016
<i>Neutral</i>						
Precision	0.51 ± 0.04	0.56 ± 0.03	0.56 ± 0.02	0.550 ± 0.038	0.534 ± 0.033	0.550 ± 0.022
Recall	0.60 ± 0.07	0.62 ± 0.05	0.58 ± 0.05	0.618 ± 0.087	0.596 ± 0.077	0.660 ± 0.038
F1	0.55 ± 0.01	0.59 ± 0.01	0.57 ± 0.04	0.578 ± 0.020	0.562 ± 0.023	0.578 ± 0.011
<i>Contradiction</i>						
Precision	0.54 ± 0.02	0.55 ± 0.05	0.53 ± 0.02	0.508 ± 0.030	0.540 ± 0.042	0.522 ± 0.021
Recall	0.504 ± 0.04	0.47 ± 0.11	0.57 ± 0.04	0.534 ± 0.133	0.486 ± 0.086	0.472 ± 0.032
F1	0.52 ± 0.02	0.51 ± 0.03	0.54 ± 0.01	0.514 ± 0.067	0.510 ± 0.032	0.492 ± 0.022

Table 5.5: Class-wise mT5-small performance (mean $\pm$ 95% CI) after synthetic(Ours) $\rightarrow$ gold training, showing annotator quality effects across NLI classes. Aya-worst vs. Aya-best, GPT-4o-mini, a DeBERTa baseline, and ensembles. Higher-quality annotators (Aya-best, GPT-4o-mini) yield consistent gains—most pronounced for *Entailment*—while *Neutral* remains comparatively difficult across models.

Metric	Aya-worst	Aya-best	GPT-4o-mini	DeBERTa-NLI	Prompts ensemble	Models ensemble
<i>Entailment</i>						
Precision	0.56 ± 0.03	0.682 ± 0.016	0.68 ± 0.03	0.656 ± 0.032	0.656 ± 0.069	0.628 ± 0.019
Recall	0.67 ± 0.07	0.574 ± 0.035	0.57 ± 0.04	0.607 ± 0.055	0.604 ± 0.131	0.634 ± 0.021
F1	0.61 ± 0.01	0.622 ± 0.019	0.62 ± 0.02	0.628 ± 0.023	0.624 ± 0.042	0.628 ± 0.016
<i>Neutral</i>						
Precision	0.60 ± 0.04	0.560 ± 0.011	0.57 ± 0.02	0.598 ± 0.027	0.570 ± 0.034	0.566 ± 0.028
Recall	0.47 ± 0.07	0.676 ± 0.019	0.64 ± 0.01	0.576 ± 0.065	0.624 ± 0.030	0.614 ± 0.031
F1	0.53 ± 0.03	0.612 ± 0.007	0.60 ± 0.01	0.582 ± 0.022	0.604 ± 0.023	0.580 ± 0.020
<i>Contradiction</i>						
Precision	0.57 ± 0.02	0.614 ± 0.032	0.59 ± 0.01	0.570 ± 0.016	0.624 ± 0.053	0.580 ± 0.023
Recall	0.56 ± 0.05	0.604 ± 0.039	0.63 ± 0.02	0.624 ± 0.028	0.596 ± 0.132	0.534 ± 0.023
F1	0.57 ± 0.01	0.608 ± 0.005	0.61 ± 0.02	0.598 ± 0.014	0.604 ± 0.042	0.554 ± 0.019

Table 5.6: Class-wise mT5-base performance (mean $\pm$ 95% CI) after synthetic(Ours) $\rightarrow$ gold training, showing annotator quality effects across NLI classes. Aya-worst vs. Aya-best, GPT-4o-mini, a DeBERTa baseline, and ensembles. Higher-quality annotators (Aya-best, GPT-4o-mini) yield consistent gains—most pronounced for *Entailment*—while *Neutral* remains comparatively difficult across models.

5.4 Discussion

5.4.1 The Dominant Effect of Annotator Quality on Downstream Performance

Our results confirm that annotator quality directly and measurably propagates through the synthetic data pipeline, impacting student model performance. The contrast between the Aya-best and Aya-worst conditions provides the most compelling evidence: a 15-point difference in annotator accuracy on the gold benchmark translates into a substantial and persistent performance gap in downstream tasks. The performance gap between Aya-best and Aya-worst is substantial in the synthetic-only and synthetic+gold settings (+0.129 and +0.121 in accuracy; +0.226 and +0.177 in weighted-F1, respectively, for mT5-base), but decreases markedly in the synthetic→gold regime (+0.043 accuracy, +0.046 weighted-F1), indicating that fine-tuning on gold data mitigates the impact of suboptimal prompting.

Notably, this quality effect was not erased by subsequent fine-tuning on gold data. While the Synthetic → Gold regime narrowed the performance gaps, models initially trained on low-quality data (e.g., Aya-worst) consistently lagged behind those trained on high-quality data (e.g., Aya-best, GPT-4o-mini). This indicates that poor-quality annotations may impart a suboptimal initial bias or representation that is not fully corrected by limited gold data, underscoring the long-term cost of using noisy synthetic labels. The class-wise variation (Figures 5.7 and 5.8) implies that annotator quality should be assessed not only by overall accuracy, but also by label distribution and class-wise reliability, as these reveal systematic biases in annotation behavior. However, our results show that these differences do not translate proportionally to downstream performance, which converges across annotator choices under the synthetic → gold training regime. This indicates that training dynamics can mitigate substantial annotation variability, highlighting a gap between annotation quality metrics and downstream effectiveness.

5.4.2 Architectural Considerations: Specialization vs. Scale

The specialized NLI model, orders of magnitude smaller, yields downstream models within 2 percentage points of GPT-4o-mini across all training regimes (synthetic-only, synthetic+gold, and synthetic→gold) on both mT5-small and mT5-base. This suggests that for well-defined tasks like NLI, task-specific proficiency can be more effective than sheer scale or general reasoning capability. The strong performance of GPT-4o-mini data confirms that state-of-the-art generalist models, when properly prompted, are highly effective annotators. However, the success of the specialized DeBERTa model offers a compelling, computationally efficient alternative for task-focused data generation, provided the data necessary to train it for the task is available.

5.4.3 The Amplifying Role of Prompt Design

The experiment comparing Aya-best and Aya-worst isolates the effect of prompt design from model capability on the downstream task. The results show that prompt engineering is not merely a peripheral concern but a fundamental component of data quality. A poorly designed prompt can induce severe label distribution skews (as seen in Table 5.1) and dramatically reduce the informational value of the generated data. Conversely, an optimized prompt can unlock a model’s annotative potential, yielding data that drives student model performance close to that achieved with top-tier proprietary models. This highlights that the annotator is not just the model, but the model-prompt system.

5.4.4 The Promise and Cost of Ensemble Methods

Our exploration of ensemble methods revealed a classic trade-off between quality and computational cost. Both inter-annotator and prompt-variant ensembles achieved higher accuracy on the gold benchmark, confirming that aggregation can mitigate individual errors and biases. The strong correlation between agreement with the ensemble and gold accuracy (Figure 5.3) further validates ensemble agreement as a reliable, unsupervised proxy for label quality. In downstream tasks, the Prompts-ensemble data consistently performed on par with the best single prompts, demonstrating that ensembling is a robust strategy for generating high-quality training data. However, these quality gains come at a high cost. The inter-annotator ensemble required $3\times$ the inference compute for a modest 2-3 point gain in accuracy. Similarly, the prompt-variant ensemble required $8\times$ the inference calls per model. This raises a critical practical question: is the marginal improvement in downstream performance justified by the substantial increase in cost? In many scenarios, using a single, well-prompted model might represent a more efficient solution. However, the design and selection of a good prompt often presupposes the existence of at least some task-specific labeled data that can be used to assess and compare the performance of different prompt alternatives. For example, our analysis in Chapter 3 used the ParsiNLU dataset to compare prompt variants. Consequently, in extremely low-resource scenarios—where there is no pre-existing labeled data for a task—our results using ensemble labeling suggest that in these cases, the use of an ensemble of prompts as an initial LLM annotator may be useful in developing a seed dataset for research that has a higher probability of generating good labels than using a single (data-blind selected/designed) prompt.

5.5 Summary

This chapter investigated the key role of label quality in LLM-generated synthetic data for NLI. Moving beyond the simple question of whether synthetic data works, we explored how the quality of the LLM annotator affects downstream utility. Our controlled experiments, which held the input text constant while varying the annotation source, reveal a clear and consistent theme: the proficiency of the LLM annotator is a primary determinant of the efficacy of synthetic data. We demonstrated this through three key lenses: model choice, prompt design, and aggregation strategies. The performance of student models trained on this data showed a clear quality gradient, from near-random performance with poorly designed prompts to competitive results with high-quality or ensemble-based annotations.



Conclusion

This chapter concludes the study by summarizing the key findings in relation to the aims and research questions. It also reviews the study's limitations and proposes opportunities for future research.

6.1 Summary of Thesis

This thesis aimed to address the critical challenge of data scarcity in low-resource languages by investigating LLMs' dual capabilities as data generators and annotators. Using Farsi NLI as a case study, we systematically evaluated the reliability of LLM-based annotations, the utility of synthetically generated data for improving downstream task performance, and the pivotal role of annotator quality. The progression of this work moved from a foundational behavioral study of LLMs as annotators (Chapter 3), to the development and evaluation of a scalable pipeline for synthetic data creation (Chapter 4), and finally to a controlled investigation of how annotator quality propagates to model performance (Chapter 5). This research provides a framework for leveraging LLMs to bridge the data gap in underrepresented languages.

6.2 Answering Research Questions

RQ1: How reliable are LLMs when used as automatic annotators in Farsi NLI? Our investigation reveals that the reliability of LLMs as automatic annotators in low-resource settings is highly conditional and multifaceted. While they possess significant annotation capabilities, their reliability is influenced by model choice, prompt design, and the task's inherent challenges. Our results on the NLI task annotation demonstrate that their performance is markedly unstable, varying with minor prompt changes, with the order of the premise and hypothesis being a critical factor. A clear model family divide also exists, with proprietary models such as Claude-Opus-4 and GPT-4o exhibiting greater robustness than more volatile open-weight models. Compounding this, we found

that perceived model unreliability sometimes reflected noise and ambiguity in the human-annotated benchmark.

RQ2: Can an LLM-based process generate synthetic data that improves the performance of NLP models on the NLI task in Farsi? Our results provide a clear affirmative answer. The synthetic data generated by our LLM-based pipeline substantially improved the performance of downstream NLI models for Farsi. We observed consistent, scalable gains: models fine-tuned on increasingly larger synthetic datasets outperformed those trained solely on the limited gold-standard data. Crucially, the strategy for integrating synthetic data was important. The most effective approach was a two-stage process of synthetic pre-training followed by gold fine-tuning (Synthetic→ Gold), which consistently achieved the highest performance across all metrics and model sizes. This regime effectively uses the large-scale synthetic data for robust initialization and the high-quality gold data for precise calibration.

RQ3: To what extent does the annotation quality of the LLM labeler affect downstream NLI task performance? Our experiments conclusively demonstrate that the quality of the LLM annotator is a dominant factor directly governing the utility of synthetic data. We established a clear performance gradient in downstream models, with data labeled by a high-performing prompt configuration yielding better results than data labeled by a low-performing one. This performance gap proved persistent, narrowing but not disappearing after subsequent fine-tuning on gold data, indicating that poor-quality annotations impart a sub-optimal bias that is not easily unlearned. While ensemble methods enhanced reliability, they also incurred substantial computational cost, creating a critical trade-off. Therefore, investing in a high-quality annotator—whether through model selection, prompt engineering, or strategic aggregation—is not merely beneficial but essential for maximizing downstream gains.

6.3 Limitations

While this research demonstrates the promise of LLM-driven data generation, it is important to acknowledge its constraints.

Data Source and Diversity. The scope of our synthetic data is inherently constrained by its source. We exclusively used the XLSum dataset, a collection of news articles, to generate premises. This introduces a domain and stylistic bias towards journalistic prose. Consequently, the linguistic diversity of our synthetic data (e.g., conversational, literary, or social media language) is limited compared to the more varied gold-standard dataset.

Models trained on this data may experience a performance drop when encountering text from domains not represented in the news-centric training distribution.

Pipeline-Inherent Biases. Our two-stage generation and labeling pipeline, while effective, exhibited a systematic classification bias, producing a heavy and persistent skew towards the Entailment class while under-generating Neutral and Contradiction examples. This imbalance is a direct result of the LLM-driven data synthesis process’s tendency to generate hypotheses that remain faithful to the premise. This fundamental bias suggests that generating challenging negative examples for NLI remains as a challenge for LLM-based data creation.

Generalizability and Specificity. Our study provides a deep, evidence-based analysis for Farsi NLI, but it remains a single-task and single-language case study. The effectiveness of our specific prompt formulations and the observations may not directly translate to other tasks and languages.

6.4 Future Work

Advanced Prompting Strategies. Future work should investigate more sophisticated labeling techniques to mitigate prompt sensitivity and improve accuracy. This includes exploring Chain-of-Thought (CoT) prompting to elicit the model’s reasoning before a final judgment. Building on our diagnostic analysis, a promising avenue is to develop knowledge-driven prompts that incorporate explicit guidelines to correct for observed failure patterns. For instance, prompts could be engineered to steer the model towards a strict semantic interpretation explicitly and to include class-specific warnings. This transforms the prompt from a generic instruction into a targeted corrective tool based on empirical behavioral findings. Another approach could be to incorporate these implications into NLI classification as additional labels [49] or to explore multi-round validation procedures to distinguish annotation errors from legitimate human variation [145].

Data Coverage, Curation, and Filtering. To enhance data coverage, future work should leverage diverse textual sources as premises. Moreover, strategic sampling methods can then be employed to select the most informative and representative instances rather than naively scaling the data size [133]. To address noise and class imbalance, future research could develop more nuanced data curation strategies. This involves creating automated quality filters based on the sensitivity and consistency metrics defined in this thesis to identify and remove ambiguous or low-confidence samples. Exploring generative

models or prompts specifically designed to generate challenging, diverse non-entailment examples would directly address the current bias towards the Entailment class [55].

Scaling to Larger Student Models. While we focused on smaller encoder-decoder models such as mT5, it is crucial to investigate the utility of our synthetic data for fine-tuning much larger, decoder-based, instruction-tuned LLMs. This would assess whether synthetic data can effectively distill task-specific knowledge into models that are closer to the *teacher* annotators themselves.

Task Diversification. The current research centered on the NLI task as a representative benchmark for semantic reasoning. However, the framework can be readily extended to other tasks that require contextual judgment or nuanced interpretation. Future work could explore task diversification by applying the proposed annotation and filtering pipeline to tasks such as sentiment analysis, question answering, summarization, or commonsense reasoning. Comparing model behavior across these task types would further illuminate whether the reliability patterns observed for NLI generalize to other forms of linguistic inference. Such extensions would also help test the general applicability of the sensitivity and consistency metrics as universal measures of annotation robustness.

Human-in-the-Loop and Active Learning Pipelines. Despite the cost efficiency of LLM annotators, their limitations in handling subjective tasks require a hybrid approach. Promising methods therefore leverage human-in-the-loop strategies to calibrate LLM performance through improved prompt design [96, 120]. Designing human-in-the-loop pipelines in which LLMs generate the bulk of the data, with human annotators strategically deployed to label the most ambiguous or high-sensitivity cases, could yield high-quality datasets more efficiently than either humans or LLMs alone.

6.5 Concluding Remarks

In conclusion, this thesis demonstrates that the strategic use of LLMs for synthetic data generation and annotation presents a powerful tool. Overall, we find that synthetic data enables performance gains that would be difficult to achieve with limited human-annotated data alone. However, this promise is tempered by significant sensitivities to prompt design and model choice. Our advice for best practice is to treat LLMs not as infallible but as powerful, yet flawed, tools whose output must be carefully managed through robust prompting, strategic training regimes, and a clear understanding of their failure modes. By doing so, we can harness their scale and capability to foster greater linguistic inclusivity and build more effective NLP systems.

Bibliography

- [1] *Morphologies of Asia and Africa*. Penn State University Press, 2007.
- [2] A. Abaskohi, S. Baruni, M. Masoudi, N. Abbasi, M. H. Babalou, A. Edalat, S. Kamahi, S. Mahdizadeh Sani, N. Naghavian, D. Namazifard, P. Sadeghi, and Y. Yaghoobzadeh. Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203, Torino, Italia, May 2024. ELRA and ICCL.
- [3] M. Aghajani, A. Badri, and H. Beigy. ParsTwiNER: A corpus for named entity recognition at informal Persian. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 131–136, Online, Nov. 2021. Association for Computational Linguistics.
- [4] R. Aharoni, S. Narayan, J. Maynez, J. Herzig, E. Clark, and M. Lapata. Multilingual summarization with factual consistency evaluation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] O. Ahia, S. Kumar, H. Gonen, J. Kasai, D. Mortensen, N. Smith, and Y. Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore, Dec. 2023. Association for Computational Linguistics.
- [6] W. Ahmad, Z. Zhang, X. Ma, E. Hovy, K.-W. Chang, and N. Peng. On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [7] K. Ahuja, S. Kumar, S. Dandapat, and M. Choudhury. Multi task learning for zero shot performance prediction of multilingual models. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland, May 2022. Association for Computational Linguistics.

- [8] H. Amirkhani, M. AzariJafari, S. Faridan-Jahromi, Z. Kouhkan, Z. Pourjafari, and A. Amirak. Farstail: a persian natural language inference dataset. *Soft Computing*, July 2023.
- [9] M. Artetxe, G. Labaka, and E. Agirre. Translation artifacts in cross-lingual transfer learning. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, Nov. 2020. Association for Computational Linguistics.
- [10] M. Artetxe, S. Ruder, and D. Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.
- [11] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.
- [12] D. Ashok and J. May. A little human data goes a long way. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 381–413, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [13] M. Y. Ayoubi, Sajjad & Davoodeh. Persianqa: a dataset for persian question answering. <https://github.com/SajjjadAyobi/PersianQA>, 2021.
- [14] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT’ 98*, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery.
- [15] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.
- [16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Nee-lakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances*

- in *Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [17] N. Calderon, R. Reichart, and R. Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [18] P. Chen, S. Yu, Z. Guo, and B. Haddow. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
- [19] X. Chen, R. A. Chi, X. Wang, and D. Zhou. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*, 2024.
- [20] C.-H. Chiang and H.-y. Lee. Can large language models be an alternative to human evaluations? In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.
- [22] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.
- [23] C. Condoravdi, D. Crouch, V. de Paiva, R. Stolle, and D. G. Bobrow. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45, 2003.
- [24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

-
- [25] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.
- [26] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer, 2005.
- [27] K. Darvishi, N. Shahbodaghkhan, Z. Abbasiantaeb, and S. Momtazi. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486, 2023.
- [28] B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing. Is GPT-3 a good data annotator? In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [29] F. Errica, D. Sanvito, G. Siracusano, and R. Bifulco. What did I do wrong? quantifying LLMs’ sensitivity and consistency to prompt engineering. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1543–1558, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics.
- [30] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847, Oct. 2021.
- [31] M. Farahani, M. Gharachorloo, and M. Manthouri. Leveraging parsbert and pre-trained mt5 for persian abstractive text summarization. In *2021 26th International computer conference, computer society of Iran (CSICC)*, pages 1–6. IEEE, 2021.
- [32] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada, July 2023. Association for Computational Linguistics.

-
- [33] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- [34] J. ForutanRad, M. HourAli, and M. a. KeyvanRad. Farsi question and answer dataset (farsiquad). *Signal and Data Processing*, 20(4), 2024.
- [35] X. Fu and W. Liu. How reliable is multilingual LLM-as-a-judge? In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [36] C. Gan and T. Mori. Sensitivity and robustness of large language models to prompt template in Japanese text classification tasks. In C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A. W. H. Zeng, B. Peng, Y. Li, and J. Li, editors, *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 1–11, Hong Kong, China, Dec. 2023. Association for Computational Linguistics.
- [37] Z. Gekhman, J. Herzig, R. Aharoni, C. Elkind, and I. Szpektor. TrueTeacher: Learning factual consistency evaluation with large language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore, Dec. 2023. Association for Computational Linguistics.
- [38] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [39] O. Ghahroodi, M. Nouri, M. V. Sanian, A. Sahebi, D. Dastgheib, E. Asgari, M. S. Baghshah, and M. H. Rohban. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *arXiv preprint arXiv:2404.06644*, 2024.
- [40] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), July 2023.
- [41] M. Glockner, V. Shwartz, and Y. Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting*

- of the *Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, 2018. Association for Computational Linguistics.
- [42] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- [43] M. Grusky, M. Naaman, and Y. Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [44] B. Guan, T. Roosta, P. Passban, and M. Rezagholizadeh. The order effect: Investigating prompt sensitivity to input order in llms. *arXiv preprint arXiv:2502.04134*, 2025.
- [45] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [46] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [47] M. U. U. Haq, D. Rigoni, and A. Sperduti. Llms as data annotators: How close are we to human performance. *arXiv preprint arXiv:2504.15022*, 2025.
- [48] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, Aug. 2021. Association for Computational Linguistics.
- [49] S. Havaldar, H. Alvari, J. Palowitch, M. J. Hosseini, S. Buthpitiya, and A. Fabrikant. Entailed between the lines: Incorporating implication into NLI. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers*), pages 32274–32290, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [50] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*, 2024.
 - [51] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, and W. Chen. AnnoLLM: Making large language models to be better crowdsourced annotators. In Y. Yang, A. Davani, A. Sil, and A. Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
 - [52] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
 - [53] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, and L. Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2022. Curran Associates Inc.
 - [54] O. Honovich, T. Scialom, O. Levy, and T. Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, 2023.
 - [55] M. J. Hosseini, A. Petrov, A. Fabrikant, and A. Louis. A synthetic data approach for domain generalization of NLI models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
 - [56] P. Hosseini, A. A. Ramaki, H. Maleki, M. Anvari, and S. A. Mirroshandel. Sentipers: a sentiment analysis corpus for persian. *arXiv preprint arXiv:1801.07737*, 2018.
 - [57] C. Huyen. *AI Engineering*. O’Reilly Media, USA, 2025.

- [58] S. Jadhav, A. Shanbhag, A. Thakurdesai, R. Sinare, and R. Joshi. On limitations of LLM as annotator for low resource languages. In M. Abbas, T. Yousef, and L. Galke, editors, *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark, Aug. 2025. Association for Computational Linguistics.
- [59] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023.
- [60] N.-J. Jiang, C. Tan, and M.-C. de Marneffe. Understanding and predicting human label variation in natural language inference through explanation. *arXiv preprint arXiv:2304.12443*, 2023.
- [61] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.
- [62] J. Kaddour and Q. Liu. Synthetic data generation in low-resource settings via fine-tuning of large language models. *arXiv preprint arXiv:2310.01119*, 2023.
- [63] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [64] O. Kashefi. Mizan: A large persian-english parallel corpus, 2020.
- [65] A. Kazemi, S. B. Natarajan Kalaivendan, J. Wagner, H. Qadeer, K. Verma, and B. Davis. Synthetic vs. gold: The role of LLM generated labels and data in cyber-bullying detection. In G. Angelova, M. Kuniolkovskaya, M. Escribe, and R. Mitkov, editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 531–540, Varna, Bulgaria, Sept. 2025. INCOMA Ltd., Shoumen, Bulgaria.
- [66] D. Khashabi, A. Cohan, S. Shakeri, P. Hosseini, P. Pezeshkpour, M. Alikhani, M. Aminnaseri, M. Bitaab, F. Brahman, S. Ghazarian, M. Gheini, A. Kabiri, R. K. Mahabadi, O. Memarrast, A. Mosallanezhad, E. Noury, S. Raji, M. S. Rasooli, S. Sadeghi, E. S. Azer, N. S. Samghabadi, M. Shafaei, S. Sheybani, A. Tazarv, and Y. Yaghoobzadeh. ParsiNLU: A suite of language understanding challenges

- for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162, 2021.
- [67] B. Kim, H. Kim, and G. Kim. Abstractive summarization of Reddit posts with multi-level memory networks. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [68] W. Kryscinski, B. McCann, C. Xiong, and R. Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346. Association for Computational Linguistics, 2020.
- [69] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In E. Blanco and W. Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics.
- [70] H. Lai, X. Liu, J. Gao, J. Cheng, Z. Qi, Y. Xu, S. Yao, D. Zhang, J. Du, Z. Hou, X. Lv, M. Huang, Y. Dong, and J. Tang. A survey of post-training scaling in large language models. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [71] V. D. Lai, N. Ngo, A. Pouran Ben Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore, Dec. 2023. Association for Computational Linguistics.
- [72] M. Laurer, W. v. Atteveldt, A. S. Casas, and K. Welbers. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*, June 2022. Publisher: Open Science Framework.
- [73] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In

- B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online, Nov. 2020. Association for Computational Linguistics.
- [74] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, and H. Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [75] J. Li, J. Chen, R. Ren, X. Cheng, X. Zhao, J.-Y. Nie, and J.-R. Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [76] Y. Li, R. Bonatti, S. Abdali, J. Wagle, and K. Koishida. Data generation using large language models for text classification: An empirical case study. *arXiv preprint arXiv:2407.12813*, 2024.
- [77] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, and Y. T. Lee. Textbooks are all you need ii: phi-1.5 technical report. September 2023.
- [78] Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma. Leveraging large language models for NLG evaluation: Advances and challenges. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
- [79] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052, 2022.
- [80] R. Liu, J. Wei, F. Liu, C. Si, Y. Zhang, J. Rao, S. Zheng, D. Peng, D. Yang, D. Zhou, et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.

- [81] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.
- [82] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, and H. Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [83] B. Ma, Y. Li, W. Zhou, Z. Gong, Y. J. Liu, K. Jasinskaja, A. Friedrich, J. Hirschberg, F. Kreuter, and B. Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [84] B. MacCartney. *Natural Language Inference*. PhD dissertation, Stanford University, Stanford, CA, 2009.
- [85] C. Masica. *The Indo-Aryan Languages*. Cambridge Language Surveys. Cambridge University Press, 1993.
- [86] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.
- [87] R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.
- [88] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve, 2023.
- [89] N. Mirzakhmedova, M. Gohsen, C. H. Chang, and B. Stein. Are large language models reliable argument quality annotators? In *Robust Argumentation Machines*:

-
- First International Conference, RATIO 2024, Bielefeld, Germany, June 5–7, 2024, Proceedings*, page 129–146, Berlin, Heidelberg, 2024. Springer-Verlag.
- [90] I. Mondshine, T. Paz-Argaman, and R. Tsarfaty. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics.
 - [91] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. L. Scao, M. S. Bari, S. Shen, Z.-X. Yong, H. Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.
 - [92] N. Nangia and S. Bowman. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4566–4575, 2019.
 - [93] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901. Association for Computational Linguistics, 2020.
 - [94] C. T. Okolo and M. Tano. Closing the gap: A call for more inclusive language technologies. *Brookings Institution*, 2023.
 - [95] OpenAI. Gpt-4 technical report, 2023.
 - [96] N. Pangakis and S. Wolken. Keeping humans in the loop: Human-centered automated annotation with generative ai. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1471–1492, 06 2025.
 - [97] S. S. ParsBench Team. Parsbench: A framework for generating high-quality persian language datasets. <https://github.com/ParsBench>, 2025. Open-source framework for Persian language dataset generation.
 - [98] J. N. Pava, C. Meinhardt, H. Badi Uz Zaman, T. Friedman, S. T. Truong, D. Zhang, E. Cryst, V. Marivate, and S. Koyejo. Mind the (Language) gap: Mapping the challenges of LLM development in low-resource language contexts. White paper, Stanford Institute for Human-Centered Artificial Intelligence (HAI) and The Asia Foundation, 2025.
 - [99] E. Pavlick and T. Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.
-

-
- [100] M. Pavlovic and M. Poesio. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia, May 2024. ELRA and ICCL.
- [101] J. R. Perry. Arabic language v. Arabic elements in Persian. In *Encyclopaedia Iranica*. 2011. Online edition, available at <https://www.iranicaonline.org/articles/arabic-v>.
- [102] B. Plank. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [103] E. M. Ponti, G. Glavaš, O. Majewska, Q. Liu, I. Vulić, and A. Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, Nov. 2020. Association for Computational Linguistics.
- [104] H. Poostchi, E. Zare Borzeshi, and M. Piccardi. BiLSTM-CRF for Persian named-entity recognition ArmanPersoNERCorpus: the first entity-annotated Persian dataset. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).
- [105] J. Pustejovsky and A. Stubbs. *Natural Language Annotation for Machine Learning*. Number v. 9, p. 878 in A Guide to corpus-building for applications. O’Reilly Media, Incorporated, 2013.
- [106] B. QasemiZadeh, S. Rahimi, and M. Safaee Ghalati. Challenges in Persian electronic text analysis. *arXiv preprint arXiv:1404.4740*, 2014.
- [107] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, and P. S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*, 2024.

- [108] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.
- [109] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [110] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics.
- [111] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.
- [112] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282, Nov. 2017.
- [113] D. Rooein, F. M. Plaza-del Arco, D. Nozza, and D. Hovy. Exploring subjective tasks in Farsi: A survey analysis and evaluation of language models. In J. Barnes, V. Barriere, O. De Clercq, R. Klinger, C. Nouri, D. Nozza, and P. Singh, editors, *The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment Social Media Analysis (WASSA 2026)*, pages 83–95, Rabat, Morocco, Mar. 2026. Association for Computational Linguistics.
- [114] A. Ross and E. Pavlick. How well do NLI models capture verb veridicality? In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.
- [115] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, Aug. 2021. Association for Computational Linguistics.

-
- [116] J. H. Rysstrøm, H. R. Kirk, and S. Hale. Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In P. Przybyła, M. Shardlow, C. Colombatto, and N. Inie, editors, *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 74–85, Varna, Bulgaria, Sept. 2025. INCOMA Ltd., Shoumen, Bulgaria.
- [117] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, T. L. Scao, S. Ruder, C. Barham, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*, 2022.
- [118] S. Santy, J. Liang, R. Le Bras, K. Reinecke, and M. Sap. NLPositionality: Characterizing design biases of datasets and models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [119] P. Schilcher, D. Karasin, M. Schöpf, H. Saleh, A. Tommasel, and M. Schedl. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20643–20664, Suzhou, China, Nov. 2025. Association for Computational Linguistics.
- [120] H. Schroeder, D. Roy, and J. Kabbara. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Annual Meeting of the Association for Computational Linguistics*, 2025.
- [121] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083, 2024.
- [122] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In K. Erk and N. A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.
- [123] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for*
-

- Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016. <https://doi.org/10.18653/v1/P16-1162>.
- [124] M. S. Shahshahani, M. Mohseni, A. Shakery, and H. Faili. Peyma: A tagged corpus for persian named entities. *ArXiv*, abs/1801.09936, 2018.
- [125] M. Shamsfard. Challenges and open problems in Persian text processing. In *Proceedings of the 5th Language and Technology Conference (LTC)*, pages 65–69, 2011.
- [126] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [127] S. Singh, F. Vargus, D. Dsouza, B. F. Karlsson, A. Mahendiran, W.-Y. Ko, H. Shandilya, J. Patel, D. Mataciunas, L. OMahony, M. Zhang, R. Hettiarachchi, J. Wilson, M. Machado, L. S. Moura, D. Krzemiński, H. Fadaei, I. Ergun, I. Okoh, A. Alaagib, O. Mudannayake, Z. Alyafeai, M. C. Vu, S. Ruder, S. Guthikonda, E. A. Alghamdi, S. Gehrmann, N. Muennighoff, M. Bartolo, J. Kreutzer, A. Ustun, M. Fadaee, and S. Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*, 2024.
- [128] R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*, 2008.
- [129] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022.
- [130] E. Taher, S. A. Hoseini, and M. Shamsfard. Beheshti-NER: Persian named entity recognition using BERT. In A. A. Freihat and M. Abbas, editors, *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 37–42, Trento, Italy, 11–12 Sept. 2019. Association for Computational Linguistics.
- [131] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu. Large language models for data annotation and synthesis: A survey. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

- [132] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023.
- [133] S. Tavakkol, M. Springer, M. Bateni, N. Bulut, V. Cohen-Addad, and M. Hajiaghayi. Less is more: adaptive coverage for synthetic training data. *arXiv preprint arXiv:2504.14508*, 2025.
- [134] S. Tehranipour. Snappfood - Persian Sentiment Analysis. <https://www.kaggle.com/datasets/soheiltehranipour/snappfood-persian-sentiment-analysis>, 2021.
- [135] F. Trad and A. Chehab. *To Ensemble or Not: Assessing Majority Voting Strategies for Phishing Detection with Large Language Models*, page 158–173. Springer Nature Switzerland, 2025.
- [136] Y.-M. Tseng, W.-L. Chen, C.-C. Chen, and H.-H. Chen. Evaluating large language models as expert annotators. *arXiv preprint arXiv:2508.07827*, 2025.
- [137] A. Üstün, V. Aryabumi, Z. Yong, W.-Y. Ko, D. D’souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, et al. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pages 15894–15939, 2024.
- [138] V. Viswanathan, X. Yue, A. Liu, Y. Wang, and G. Neubig. Synthetic data in the era of large language models. In Y. Arase, D. Jurgens, and F. Xia, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 11–12, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [139] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [140] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

- [141] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, and Z. Sui. Large language models are not fair evaluators. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [142] S. Wang, H. Fang, M. Khabsa, H. Mao, and H. Ma. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*, 2021.
- [143] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.
- [144] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [145] L. Weber-Genzel, S. Peng, M.-C. De Marneffe, and B. Plank. VariErr NLI: Separating annotation error from human label variation. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.
- [146] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.
- [147] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.
- [148] J. Wei and K. Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

-
- [149] C. Whitehouse, M. Choudhury, and A. F. Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore, Dec. 2023. Association for Computational Linguistics.
- [150] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [151] B. Workshop, :, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurençon, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klammer, C. Leong, D. van Strien, D. I. Adelani, D. Radev, E. G. Ponferrada, E. Levkovich, E. Kim, E. B. Natan, F. D. Toni, G. Dupont, G. Kruszewski, G. Pistilli, H. El-sahar, H. Benyamina, H. Tran, I. Yu, I. Abdulmumin, I. Johnson, I. Gonzalez-Dios, J. de la Rosa, J. Chim, J. Dodge, J. Zhu, J. Chang, J. Frohberg, J. Tobing, J. Bhattacharjee, K. Almubarak, K. Chen, K. Lo, L. V. Werra, L. Weber, L. Phan, L. B. allal, L. Tanguy, M. Dey, M. R. Muñoz, M. Masoud, M. Grandury, M. Šaško, M. Huang, M. Coavoux, M. Singh, M. T.-J. Jiang, M. C. Vu, M. A. Jauhar, M. Ghaleb, N. Subramani, N. Kassner, N. Khamis, O. Nguyen, O. Espejel, O. de Gibert, P. Villegas, P. Henderson, P. Colombo, P. Amuok, Q. Lhoest, R. Harliman, R. Bommasani, R. L. López, R. Ribeiro, S. Osei, S. Pyysalo, S. Nagel, S. Bose, S. H. Muhammad, S. Sharma, S. Longpre, S. Nikpoor, S. Silberberg, S. Pai, S. Zink, T. T. Torrent, T. Schick, T. Thrush, V. Danchev, V. Nikoulina, V. Laipala, V. Lepercq, V. Prabhu, Z. Alyafeai, Z. Talat, A. Raja, B. Heinzerling, C. Si, D. E. Taşar, E. Salesky, S. J. Mielke, W. Y. Lee, A. Sharma, A. Santilli, A. Chaffin, A. Stiegler, D. Datta, E. Szczechla, G. Chhablani, H. Wang, H. Pandey, H. Strobelt, J. A. Fries, J. Rozen, L. Gao, L. Sutawika, M. S. Bari, M. S. Al-shaibani, M. Manica, N. Nayak, R. Teehan, S. Albanie, S. Shen, S. Ben-David, S. H. Bach, T. Kim, T. Bers, T. Fevry, T. Neeraj, U. Thakker, V. Raunak, X. Tang, Z.-X. Yong, Z. Sun, S. Brody, Y. Uri, H. Tojarieh, A. Roberts, H. W. Chung, J. Tae, J. Phang,
-

O. Press, C. Li, D. Narayanan, H. Bourfoune, J. Casper, J. Rasley, M. Ryabinin, M. Mishra, M. Zhang, M. Shoeybi, M. Peyrounette, N. Patry, N. Tazi, O. Sanseviero, P. von Platen, P. Cornette, P. F. Lavallée, R. Lacroix, S. Rajbhandari, S. Gandhi, S. Smith, S. Requena, S. Patil, T. Dettmers, A. Baruwa, A. Singh, A. Cheveleva, A.-L. Ligozat, A. Subramonian, A. Névél, C. Lovering, D. Garrette, D. Tunuguntla, E. Reiter, E. Taktasheva, E. Voloshina, E. Bogdanov, G. I. Winata, H. Schoelkopf, J.-C. Kalo, J. Novikova, J. Z. Forde, J. Clive, J. Kasai, K. Kawamura, L. Hazan, M. Carpuat, M. Clinciu, N. Kim, N. Cheng, O. Serikov, O. Antverg, O. van der Wal, R. Zhang, R. Zhang, S. Gehrmann, S. Mirkin, S. Pais, T. Shavrina, T. Scialom, T. Yun, T. Limisiewicz, V. Rieser, V. Protasov, V. Mikhailov, Y. Pruksachatkun, Y. Belinkov, Z. Bamberger, Z. Kasner, A. Rueda, A. Pestana, A. Feizpour, A. Khan, A. Faranak, A. Santos, A. Hevia, A. Undreaj, A. Aghagol, A. Abdollahi, A. Tammour, A. HajiHosseini, B. Behroozi, B. Ajibade, B. Saxena, C. M. Ferrandis, D. McDuff, D. Contractor, D. Lansky, D. David, D. Kiela, D. A. Nguyen, E. Tan, E. Baylor, E. Ozoani, F. Mirza, F. Ononiwu, H. Rezanejad, H. Jones, I. Bhattacharya, I. Solaiman, I. Sedenko, I. Nejadgholi, J. Passmore, J. Seltzer, J. B. Sanz, L. Dutra, M. Samagaio, M. Elbadri, M. Mieskes, M. Gerchick, M. Akinlolu, M. McKenna, M. Qiu, M. Ghauri, M. Burynek, N. Abrar, N. Rajani, N. Elkott, N. Fahmy, O. Samuel, R. An, R. Kromann, R. Hao, S. Alizadeh, S. Shubber, S. Wang, S. Roy, S. Viguier, T. Le, T. Oyebade, T. Le, Y. Yang, Z. Nguyen, A. R. Kashyap, A. Palasciano, A. Callahan, A. Shukla, A. Miranda-Escalada, A. Singh, B. Beilharz, B. Wang, C. Brito, C. Zhou, C. Jain, C. Xu, C. Fourrier, D. L. Periñán, D. Molano, D. Yu, E. Manjavacas, F. Barth, F. Fuhrmann, G. Altay, G. Bayrak, G. Burns, H. U. Vrabec, I. Bello, I. Dash, J. Kang, J. Giorgi, J. Golde, J. D. Posada, K. R. Sivaraman, L. Bulchandani, L. Liu, L. Shinzato, M. H. de Bykhovetz, M. Takeuchi, M. Pàmies, M. A. Castillo, M. Nezhurina, M. Sängler, M. Samwald, M. Cullan, M. Weinberg, M. D. Wolf, M. Mihaljcic, M. Liu, M. Freidank, M. Kang, N. Seelam, N. Dahlberg, N. M. Broad, N. Muellner, P. Fung, P. Haller, R. Chandrasekhar, R. Eisenberg, R. Martin, R. Canalli, R. Su, R. Su, S. Cahyawijaya, S. Garda, S. S. Deshmukh, S. Mishra, S. Kiblawi, S. Ott, S. Sangaroonsiri, S. Kumar, S. Schweter, S. Bharati, T. Laud, T. Gigant, T. Kainuma, W. Kusa, Y. Labrak, Y. S. Bajaj, Y. Venkatraman, Y. Xu, Y. Xu, Y. Xu, Z. Tan, Z. Xie, Z. Ye, M. Bras, Y. Belkada, and T. Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023.

- [152] S. Wu and M. Dredze. Are all languages created equal in multilingual bert? In S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, and H. Hajishirzi, editors, *Proceedings of the 5th Workshop on Representation Learn-*

- ing for NLP*, pages 120–130. Association for Computational Linguistics, 2020. <https://doi.org/10.18653/v1/2020.repl4nlp-1.16>.
- [153] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698, 2020.
- [154] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.
- [155] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.
- [156] M. Zarharan, M. Hashemi, M. Behroozrazegh, S. Eetemadi, M. T. Pilehvar, and J. Foster. FarExStance: Explainable stance detection for Farsi. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10125–10147. Association for Computational Linguistics, 2025.
- [157] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- [158] J. Zhuo, S. Zhang, X. Fang, H. Duan, D. Lin, and K. Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.