Evaluating Trustworthy Multilingual Alignment in LLMs via Synthetic Natural Language Inference

Samir Abdaljalil[1,*], Erchin Serpedin[1], Khalid Qaraqe[2] and Hasan Kurban[2]

[1]Texas A&M University, College Station TX, USA
[2]Hamad Bin Khalifa University, Doha, Qatar

Abstract

Large language models (LLMs) are increasingly deployed in multilingual settings, where inconsistent behavior across languages can undermine reliability, fairness, and user trust. We present a controlled evaluation framework for multilingual natural language inference (NLI) that stress-tests semantic consistency using synthetic logic-based premise–hypothesis pairs in both monolingual and code-switched conditions. Because several models perform near the 33.3% random baseline for the balanced three-class task, we frame the benchmark as a diagnostic stress test of multilingual reasoning brittleness rather than evidence of robust NLI performance. In some model–language configurations, code-switched inputs match or exceed monolingual performance; however, these effects are uneven and may reflect prompting effects, translation artifacts, answer-extraction sensitivity, or label-calibration differences rather than genuine reasoning gains. Embedding-based similarity and cross-lingual alignment analyses provide evidence that translated pairs generally preserve semantic content. Overall, our findings expose the brittleness of current LLM cross-lingual reasoning and show that code-switching is useful as a diagnostic condition for multilingual robustness evaluation. Code is available at: https://github.com/KurbanIntelligenceLab/nli-stress-testing.

Keywords

Large Language Models (LLMs), Natural Language Inference (NLI), Multilingual Alignment

1. Introduction

NLI [1]—determining whether a *hypothesis* is entailed by, contradicts, or is neutral with respect to a *premise*—is a core benchmark for natural language understanding [2, 3, 4]. Its emphasis on fine-grained semantic distinctions has long made it a proxy for testing models' capacity for deep reasoning [5]. With LLMs, NLI has become a key tool for assessing generalization, reasoning, and knowledge encoding [6]. Yet evaluations remain concentrated on high-resource languages—especially English—and are often embedded within downstream tasks such as QA or summarization, limiting insight into whether inference capabilities transfer consistently across languages under controlled semantic conditions. From a trustworthy AI perspective, multilingual inconsistency presents a critical deployment challenge [7]. Models that behave differently across languages may introduce reliability failures, unequal user experiences, and hidden biases that affect multilingual populations.

We address this gap with a synthetic multilingual NLI framework that stress-tests cross-lingual semantic alignment using deterministic, logic-based templates encoding entailment, contradiction,

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.

*Corresponding author.
sabdaljalil@tamu.edu (S. Abdaljalil); eserpedin@tamu.edu (E. Serpedin); kqaraqe@hbku.edu.qa (K. Qaraqe); hkurban@hbku.edu.qa (H. Kurban)

and neutrality. The framework decouples logical structure from lexical and cultural priors, reducing annotation noise and enabling controlled large-scale evaluation. Our contributions are: (1) a logic-driven method for generating synthetic multilingual NLI datasets; (2) an automated protocol for evaluating cross-lingual consistency in LLM semantic judgments; and (3) empirical evidence that multilingual and code-switched NLI performance varies substantially across models, languages, and label types.

By disentangling logical reasoning from linguistic noise, our framework offers a principled, reproducible basis for evaluating semantic alignment in multilingual LLMs. Section 2 reviews related work, Section 3 details the methodology, and Section 4 outlines the experimental setup. Section 5 reports the main findings, followed by qualitative and quantitative analyses in Section 6. Section 7 concludes with a discussion of limitations and future directions.

2. Related Work

Natural Language Inference for Multilingual Evaluation. NLI has become a standard probe for semantic understanding in language models [8]. By requiring systems to determine whether a hypothesis follows from a premise, it offers a fine-grained test of reasoning, world knowledge, and linguistic nuance. Benchmarks such as GLUE [9] and SNLI [10] established its role in English-centric NLP, while XNLI [11] extended evaluation to 15+ languages via professional translation. Owing to its structured and interpretable format, NLI has been widely used for assessing cross-lingual transfer [12, 13]. However, most prior work assumes monolingual evaluation—premise and hypothesis in the same language—thus overlooking mixed-lingual scenarios that are common in real multilingual discourse.

Cross-Lingual Generalization in Large Language Models. Multilingual LLMs exhibit strong zero-shot transfer across languages [14, 15], aided by shared tokenization schemes and aligned embedding spaces. Early work with mBERT and XLM-R demonstrated cross-lingual transfer without explicit parallel training, attributed to emergent language alignment [16]. However, later studies revealed systematic biases: performance favors high-resource languages, while low-resource and morphologically rich languages often show degraded representations [17]. Although recent benchmarks broaden multilingual evaluation, they typically assume monolingual inputs or perfect translation symmetry. Robustness in mixed-lingual settings—where premise and hypothesis are in different languages—remains largely untested, despite its relevance for assessing sentence-level semantic alignment beyond token overlap. Code-switching, a natural phenomenon in multilingual communities, is particularly underexplored in LLM reasoning tasks [18]. Moreover, most studies use natural text, conflating syntactic variation with semantic difficulty.

Our work follows the tradition of NLI as a diagnostic tool but diverges in three ways: we use fully synthetic, logically controlled data; we evaluate translation consistency alongside reasoning; and we incorporate code-switching to probe multilingual alignment under conditions rarely addressed in prior studies. We address this by evaluating synthetic NLI pairs with controlled logical structure, enabling isolation of semantic consistency from linguistic noise. Our framework combines synthetic NLI data, high-quality translation, and controlled code-switching to stress-test multilingual alignment in both monolingual and mixed-lingual conditions. This design uncovers unexpected generalization patterns in instruction-tuned LLMs, challenging prevailing assumptions about cross-lingual reasoning robustness.

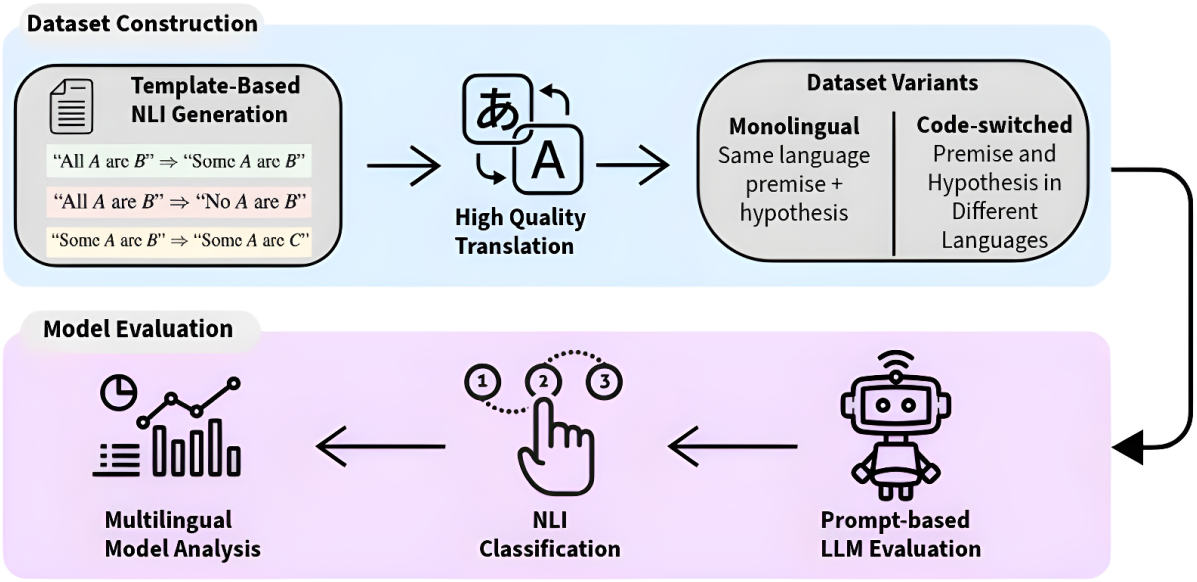

Figure 1: Pipeline for Multilingual NLI Creation and Evaluation: (1) generating NLI examples using logic-based templates, (2) translating into multiple languages with high-quality translation, (3) creating dataset variants in monolingual and code-switched formats, (4) evaluating with prompt-based LLM classification, and (5) analyzing multilingual model performance.

3. Methodology

This study examines the ability of LLMs to perform logically grounded NLI across languages using a controlled framework based on synthetic data generation and high-quality translation. Figure 1 illustrates the overall methodology for dataset construction and LLM evaluation.

Synthetic NLI Construction. A synthetic English NLI dataset is constructed from hand-crafted templates encoding three logical relations: entailment, contradiction, and neutrality. Each premise–hypothesis pair is derived from abstract quantifier-based patterns, with placeholders A, B, and C populated using semantically coherent noun phrases to ensure plausibility. The template-based design affords precise control over compositional structure and minimizes linguistic noise, thereby isolating reasoning ability from lexical variation. Figure 2 presents the templates alongside example instances from the dataset.

Logical assumptions: Labels are assigned under a closed-template interpretation in which the relation is determined only by the stated premise and hypothesis, not by external world knowledge. Entailment and contradiction examples are generated from explicit quantifier relations such as all, some, and no. Neutral examples are constructed so that the hypothesis introduces a property or category not licensed by the premise. We avoid examples where existential assumptions or real-world knowledge could plausibly change the intended label.

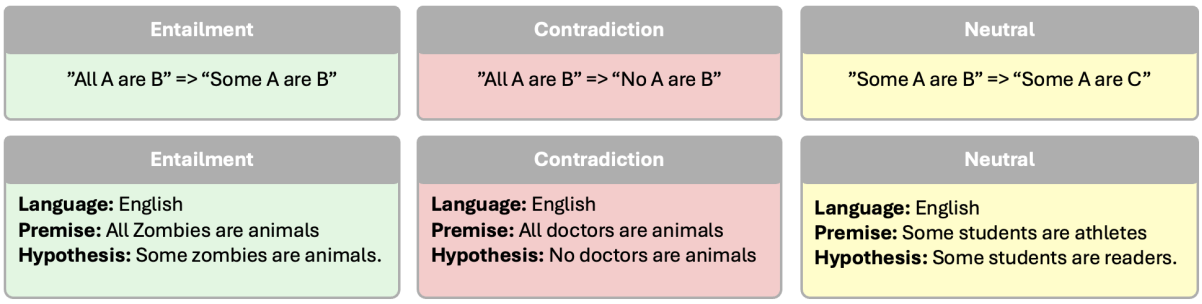

Figure 2: Top row: Synthetic NLI templates encoding entailment, contradiction, and neutrality. Placeholders *A*, *B*, and *C* are later instantiated with semantically coherent noun phrases. **Bottom row:** Samples from the generated NLI dataset for English (en), each showing one of the three relationships: entailment (green), contradiction (red), and neutral (yellow).

Multilingual Translation. To assess inference consistency across languages, the English dataset is translated into a typologically and script-diverse set of target languages[1]. The selected languages—Arabic (ar), German (de), French (fr), Hindi (hi), and Swahili (sw)—cover both high- and low-resource settings and span multiple language families: Afro-Asiatic, Indo-European (Germanic, Romance, and Indic branches), and Niger-Congo. Their scripts (Latin, Arabic, Devanagari) and varying morphological complexity, syntactic structure, and resource availability provide a comprehensive basis for evaluating model robustness and cross-lingual generalization, surfacing weaknesses that might remain hidden in high-resource-only evaluations.

Code-Switching Robustness Evaluation. To further stress-test semantic alignment, a code-switching condition is introduced in which the premise and hypothesis are presented in different languages. For each ordered pair of languages L_1 and L_2, examples are constructed with the premise in L_1 and the hypothesis in L_2, covering all possible combinations within the selected language set. This setup evaluates whether models can preserve semantic accuracy under mixed-lingual input—a common phenomenon in multilingual communication yet rarely assessed in a controlled, systematic manner.

Model Evaluation. Model behavior is assessed using a prompt-based classification setup. For each example, the LLM receives a structured prompt of the form:

NLI Prompt Example

Premise: *[premise]*
Hypothesis: *[hypothesis]*
Question: Is the hypothesis entailed by the premise, contradicted by it, or unrelated?
Answer with one of: Entailment, Contradiction, Neutral.
Answer:

[1]Translations were produced automatically using the Helsinki-NLP/opus-mt-* model series via the Hugging Face Transformers library

Output normalization: Because instruction-tuned LLMs may produce labels with additional text or minor formatting variation, model outputs are normalized before evaluation. We lowercase outputs, remove punctuation and surrounding whitespace, and map common variants to the three valid labels. Variants such as "entailed," "entails," and "entailment" are mapped to Entailment; "contradicted," "contradicts," and "contradiction" are mapped to Contradiction; and "neutral," "unrelated," and "neither" are mapped to Neutral. Outputs that cannot be mapped unambiguously are counted as invalid and treated as incorrect. We report invalid-output rates separately to distinguish answer-formatting failures from label-selection errors.

The model is instructed to output one of the three categorical labels, and the generated response is parsed using the normalization procedure above. Low-temperature decoding is applied to reduce generation variability. Predictions are evaluated against gold-standard labels, and accuracy is computed across all languages and code-switching configurations.

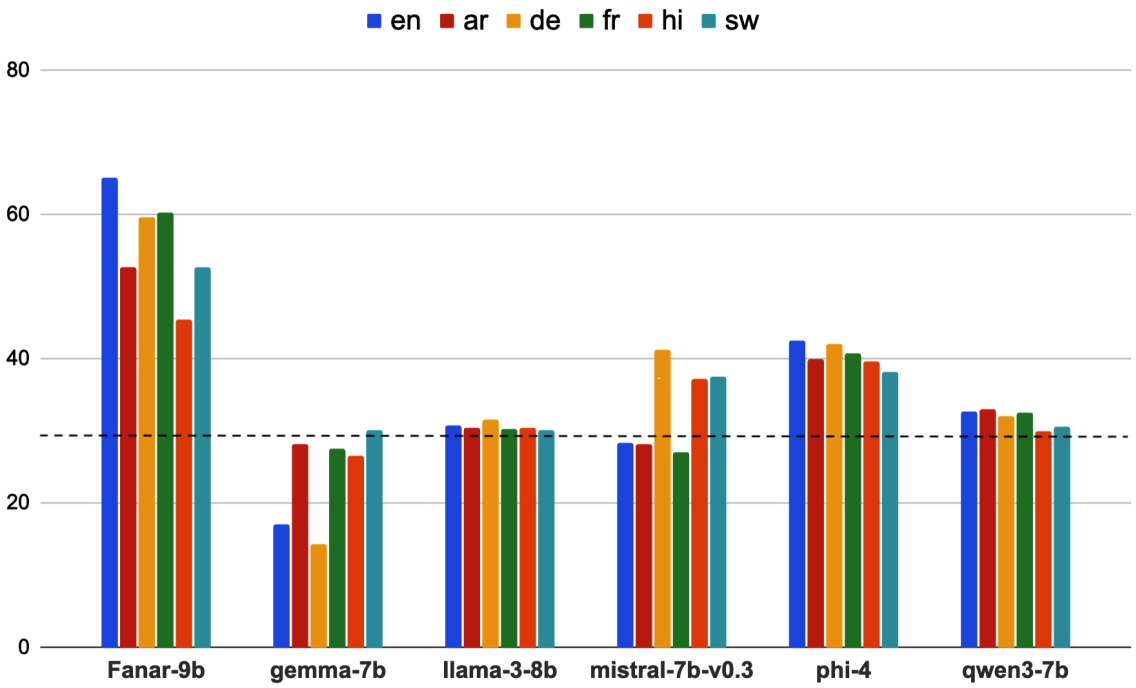

Figure 3: Monolingual NLI accuracy across six languages: English (En), Arabic (Ar), German (De), French (Fr), Hindi (Hi), and Swahili (Sw); and six LLMs. Each bar represents the accuracy of an LLM when performing natural language inference on examples where both the premise and hypothesis are in the same language. The dashed horizontal line marks the 33.3% random baseline for the balanced three-class task.

4. Experiments

Implementation Details. All experiments are executed using the Hugging Face Transformers library with a PyTorch backend. Inference is performed on A100 GPUs with `device_map="auto"` enabled for

memory-efficient model parallelism. Generation uses decoding with a maximum of 10 new tokens per prompt to produce concise outputs while limiting hallucinations, with temperature fixed at 1.0. All models are evaluated in a zero-shot setting without task-specific fine-tuning.

Models Evaluated. Six multilingual instruction-tuned LLMs are evaluated, selected for diversity in architecture, size, and training data. The set includes Fanar-9B [19], a multilingual model optimized for typologically diverse inputs; Gemma-7B [20], a decoder-only Transformer released in an instruction-tuned variant; LLaMA-3-8B [21], Meta's third-generation open-weight model pretrained on a multilingual corpus; Mistral-7B-v0.3 [22], a compact model with broad multilingual coverage; Phi-4 [23], a small but capable instruction-tuned model with strong zero-shot reasoning for its size; and Qwen3-7B [24], a multilingual model trained with extensive Chinese and non-English content. All models are evaluated using the same structured prompt format across all examples and languages to ensure comparability.

Evaluation Scope. The evaluation covers 36 language pairings (6×6) with 1,000 examples per pairing, balanced across the three NLI labels: Entailment, Contradiction, and Neutral. Both monolingual and code-switched configurations (Section 3) are included. Performance is reported using normalized label matching between parsed model predictions and gold labels. In addition to accuracy, we report invalid-output rates, class-wise precision, recall, F1, macro-F1, and confusion matrices. Since the dataset is balanced across three labels, random-chance accuracy is 33.3%, which we use as a baseline when interpreting results.

Reproducibility. All experiments use publicly available model weights and reproducible scripts. The complete setup, including prompt formatting, dataset construction, translation, and inference, is implemented in Python, enabling replication and extension to additional languages and models.

5. Results

5.1. Main Results

Monolingual inference accuracy is evaluated across six languages. In this setting, both the premise and hypothesis are in the same language, providing a baseline measure of each model's semantic reasoning capacity without cross-lingual interference. Results for the six evaluated LLMs are shown in Figure 3.

Because the dataset is balanced across three labels, 33.3% accuracy represents the random-chance baseline. Scores near this value should therefore be interpreted cautiously. In this work, near-chance performance is not treated as evidence of successful reasoning; rather, it indicates brittleness in multilingual NLI under controlled logical conditions. The diagnostic value of the benchmark lies in revealing when models collapse toward particular labels, fail to preserve semantic judgments across languages, or behave differently under code-switched inputs.

Overall Trends. Fanar-9B achieves the highest accuracy among the evaluated models, reaching 65% in English and maintaining comparatively higher performance in Swahili and Hindi. However, even for the strongest model, performance is far from saturated, indicating that the task remains challenging. These results suggest that Fanar-9B is better calibrated for this multilingual NLI setting than the other evaluated models, but they should not be interpreted as evidence of fully robust cross-lingual logical reasoning. In contrast, Gemma-7B records the lowest accuracy across nearly all languages, including 17% in English and 14% in German. The gap between the two models exceeds 40 percentage points in English, underscoring substantial differences in multilingual reasoning quality across model families.

Fanar-9b

Prem.	En	Ar	De	Fr	Hi	S
En	–	51.7	52.4	59.4	44.6	43
Ar	57.8	–	61.4	64.3	47.2	50
De	53.8	49.1	–	67.0	47.7	46
Fr	60.6	51.7	63.5	–	44.7	44
Hi	47.4	48.5	43.6	48.3	–	44
Sw	48.8	51.6	41.3	54.1	45.6	-

Gemma-7b

Prem.	En	Ar	De	Fr	Hi	S
En	–	29.4	26.2	24.8	32.9	26
Ar	40.8	–	39.3	38.8	37.4	26
De	30.1	36.6	–	29.7	36.7	24
Fr	33.5	40.7	34.7	–	40.1	25
Hi	42.6	42.2	43.2	39.4	–	25
Sw	40.0	33.6	35.9	41.7	28.4	-

Llama-3-8b

Prem.	En	Ar	De	Fr	Hi	Sw
En	–	30.0	32.0	31.0	29.0	30.0
Ar	30.8	–	31.2	30.5	30.6	29.5
De	31.7	30.2	–	29.4	30.0	29.6
Fr	30.6	30.9	30.5	–	30.3	31.1
Hi	31.1	31.2	31.4	32.7	–	30.9
Sw	32.1	32.1	31.0	31.1	31.5	–

Mistral-7b-v0.3

Prem.	En	Ar	De	Fr	Hi	S
En	–	37.0	37.7	31.0	32.4	35
Ar	36.4	–	35.7	34.8	33.5	34
De	42.2	37.6	–	36.4	35.1	36
Fr	34.5	36.4	32.9	–	34.7	34
Hi	32.9	34.5	34.9	29.5	–	35
Sw	32.5	32.1	32.4	30.3	31.9	-

Phi-4

Prem.	En	Ar	De	Fr	Hi	S
En	–	38.5	38.4	42.2	36.6	32
Ar	37.3	–	38.3	39.4	38.1	35
De	38.2	38.3	–	40.9	38.5	32
Fr	38.0	38.3	38.6	–	35.9	32
Hi	33.3	36.8	36.7	37.1	–	33
Sw	35.4	36.4	34.3	35.8	33.4	-

Qwen3-7b

Prem.	En	Ar	De	Fr	Hi	Sw
En	–	32.5	31.7	32.0	29.3	30.4
Ar	32.6	–	31.7	31.4	30.6	31.2
De	34.3	31.8	–	30.8	30.0	30.1
Fr	32.8	31.1	31.7	–	30.1	31.6
Hi	30.4	31.7	30.2	32.3	–	31.0
Sw	32.8	32.5	32.6	31.5	30.9	–

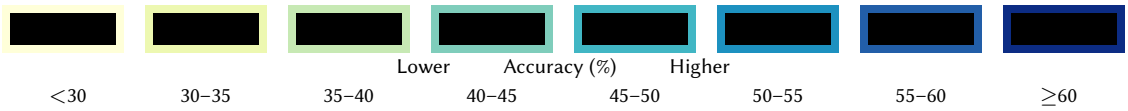

Table 1
Pairwise cross-lingual NLI accuracies (%) across six models and six languages. Rows indicate premise language and columns indicate hypothesis language; diagonal grey cells mark monolingual settings, and off-diagonal cells show code-switched performance. Colors indicate accuracy from low to high. Since labels are balanced, 33.3% is random chance; values near this baseline indicate limited reliable NLI ability.

Language-Specific Patterns. Across models, English does not consistently achieve the highest monolingual accuracy across models. In several cases, other languages match or exceed it—most notably in Mistral, where German (41%) substantially outperforms English (29%), and in Gemma-7B, where Arabic, French, and Swahili all exceed English (17%). Phi-4 performs comparably across English (43%), Arabic (40%), and German (41%), while LLaMA-3-8B shows minimal variance with scores clustered near 30% across all languages. These patterns suggest that monolingual accuracy is shaped more by model-specific pretraining distributions than by a universal high-resource advantage for English. Notably, Swahili does not consistently underperform despite its lower-resource status—in Fanar-9B and Gemma-7B, Swahili accuracy is comparable to that of Indo-European languages, likely reflecting expanded low-resource coverage in recent pretraining pipelines.

The results reveal substantial variation in monolingual reasoning performance across both languages and model architectures. Model size alone is not a reliable predictor of performance: LLaMA-3-8B underperforms relative to the smaller Phi-4, suggesting that training data composition, multilingual coverage, and instruction tuning play a more decisive role than parameter count in shaping cross-lingual logical generalization.

5.2. Code-switching

We evaluate code-switching by placing the premise and hypothesis in different languages. Table 1 reports accuracy across all language pairs, with off-diagonal cells representing bilingual inference. This setup tests whether models maintain logical consistency under mixed-language inputs.

Uneven Effects of Code-Switching. Several models outperform their monolingual baselines in specific code-switched settings. For example, Gemma-7B improves on some bilingual pairs relative to English–English, and Mistral-7B-v0.3 performs better on selected cross-lingual inputs than on some monolingual cases. However, these gains are uneven across models and language pairs. We therefore interpret them as evidence that code-switching changes model behavior, not that it generally improves reasoning.

Model- and Language-Specific Patterns. Fanar-9B achieves the highest overall accuracy across monolingual and cross-lingual settings. In contrast, Gemma-7B and Qwen3-7B show strong asymmetries: despite weak English performance, accuracy often improves when the hypothesis is in a non-English language. In several models, Hindi, Swahili, or Arabic hypotheses yield higher accuracy than English, possibly due to translation-induced simplification, prompt sensitivity, label-calibration effects, or language-specific surface forms. This aligns with prior findings that models may overfit high-resource language artifacts while benefiting from simpler translations [25].

Overall, code-switched inputs sometimes match or exceed monolingual performance, but the effect is not uniform. We therefore treat code-switching as a diagnostic condition for exposing cross-lingual asymmetries and model-specific failure modes rather than as a general strategy for improving reasoning.

6. Cross-Lingual Analysis

6.1. Embedding Similarity Across Translations

Semantic preservation across translations is examined by visualizing sentence embeddings for five randomly selected English premise statements and their translations into six languages. Sentences are encoded with LaBSE [26] into high-dimensional vectors, then projected into three dimensions using UMAP for interpretability.

Cross-Lingual Cohesion and Variation. Fig. 4 shows that translations of the same sentence generally form tight clusters across languages, suggesting that semantic meaning is largely preserved despite differences in word order, morphology, and script. Some languages show mild drift from sentence centroids; for example, Swahili deviates slightly, likely due to typological or morphological differences, consistent with prior findings on multilingual embedding spaces [27].

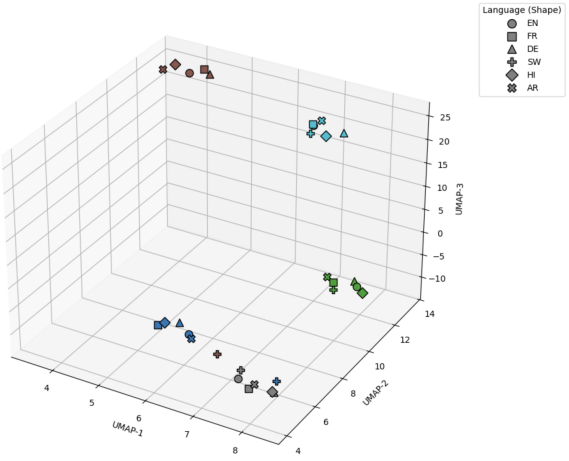

Figure 4: 3D UMAP projection of sentence embeddings across six languages. Each point represents a translation of one of five randomly selected NLI premises.

6.2. Translation Quality Assessment

We assess translation quality by computing LaBSE cosine similarity between each English sentence and its translated counterpart. As shown in Table 2, all languages achieve high average similarity, including lower-resource languages like Swahili, above 0.8, suggesting that semantic content is generally preserved. Nevertheless, because automatic translation may still affect quantifiers, negation, and syntactic scope, we treat translation quality as a remaining limitation rather than a fully eliminated confound.

Language	Code	Avg. Cosine Similarity
French	fr	0.912
German	de	0.895
Swahili	sw	0.841
Hindi	hi	0.828
Arabic	ar	0.811

Table 2: Semantic similarity between English premises and their translations. Darker blue indicates higher similarity.

7. Conclusion

This study provides a controlled evaluation of multilingual semantic alignment in instruction-tuned LLMs through a synthetic, logic-based NLI framework incorporating translation and code-switching. The design enables controlled comparison across languages and scripts while reducing confounding linguistic noise. Results show that code-switched settings can sometimes match or exceed monolingual performance, but these effects are uneven and model-dependent. Rather than demonstrating general robustness, the findings show that multilingual NLI behavior is sensitive to language pairing, prompt interpretation, answer calibration, and translation effects.

Embedding analyses reveal strong interlingual clustering of semantically equivalent sentences, supporting the feasibility of controlled multilingual evaluation. At the same time, the near-chance performance of several models highlights the brittleness of current LLMs on multilingual logical inference. Overall, the framework enables fine-grained probing of cross-lingual logic, language-specific artifacts, and code-switching as a diagnostic condition for multilingual NLP evaluation. These findings highlight both the challenges and opportunities for advancing reasoning-oriented multilingual evaluation.

Limitations. The dataset gives us control over the logical structure of the examples, but it is still synthetic and may not capture the ambiguity or variation of naturally occurring text. Several models also score close to the 33.3% random baseline, so the results should be read mainly as evidence of where multilingual NLI breaks down, rather than as proof of robust reasoning. Another limitation is the reliance on machine translation, especially because small changes in quantifiers, negation, or scope can alter the intended NLI label. Finally, because the analysis is based mostly on aggregate accuracy, apparent code-switching gains should be treated cautiously; they may reflect translation simplification, prompt sensitivity, label preferences, or calibration effects rather than genuine reasoning improvements.

Declaration on Generative AI

The authors used GPT-5 for Grammar and spelling check. After using these tool(s), the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. Dagan, O. Glickman, B. Magnini, The pascal recognising textual entailment challenge, MLCW'05, Springer-Verlag, Berlin, Heidelberg, 2005, p. 177–190. URL: https://doi.org/10.1007/11736790_9. doi:10.1007/11736790_9.

[2] S. Havaldar, H. Alvari, J. Palowitch, M. J. Hosseini, S. Buthpitiya, A. Fabrikant, Entailed between the lines: Incorporating implication into nli, 2025. URL: https://arxiv.org/abs/2501.07719. arXiv:2501.07719.

[3] F. Yudanto, Y. Sari, M. Z. A. C. Zaki, Climate-NLI: A model for natural language inference and zero-shot classification on climate-related text, in: N. Oco, S. N. Dita, A. M. Borlongan, J.-B. Kim (Eds.), Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, Tokyo University of Foreign Studies, Tokyo, Japan, 2024, pp. 600–608. URL: https://aclanthology.org/2024.paclic-1.57/.

[4] G. Mor-Lan, E. Levi, Exploring factual entailment with NLI: A news media study, in: D. Bollegala, V. Shwartz (Eds.), Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 190–199. URL: https://aclanthology.org/2024.starsem-1.15/. doi:10.18653/v1/2024.starsem-1.15.

[5] A. Cosma, S. Ruseti, M. Dascalu, C. Caragea, How hard is this test set? NLI characterization by exploiting training dynamics, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 2990–3001. URL: https://aclanthology.org/2024.emnlp-main.175/. doi:10.18653/v1/2024.emnlp-main.175.

[6] L. Cheng, T. Li, Z. Wang, T. Liu, M. Steedman, Neutralizing bias in LLM reasoning using entailment graphs, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 13714–13730. URL: https://aclanthology.org/2025.findings-acl.705/. doi:10.18653/v1/2025.findings-acl.705.

[7] V. Briva-Iglesias, Artificial intelligence language technologies in multilingual healthcare: Grand challenges ahead, 2026. URL: https://arxiv.org/abs/2605.01441. arXiv:2605.01441.

[8] A. Nighojkar, A. Laverghetta Jr., J. Licato, No strong feelings one way or another: Re-operationalizing neutrality in natural language inference, in: J. Prange, A. Friedrich (Eds.), Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 199–210. URL: https://aclanthology.org/2023.law-1.20/. doi:10.18653/v1/2023.law-1.20.

[9] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: T. Linzen, G. Chrupała, A. Alishahi (Eds.), Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: https://aclanthology.org/W18-5446/. doi:10.18653/v1/W18-5446.

[10] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 632–642. URL: https://aclanthology.org/D15-1075/. doi:10.18653/v1/D15-1075.

[11] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, V. Stoyanov, XNLI:

Evaluating cross-lingual sentence representations, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2475–2485. URL: https://aclanthology.org/D18-1269/. doi:10.18653/v1/D18-1269.

[12] M. Heredia, J. Etxaniz, M. Zulaika, X. Saralegi, J. Barnes, A. Soroa, XNLIeu: a dataset for cross-lingual NLI in Basque, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4177–4188. URL: https://aclanthology.org/2024.naacl-long.234/. doi:10.18653/v1/2024.naacl-long.234.

[13] D. Bandyopadhyay, A. De, B. Gain, T. Saikh, A. Ekbal, A deep transfer learning method for cross-lingual natural language inference, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3084–3092. URL: https://aclanthology.org/2022.lrec-1.330/.

[14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[15] M. Artetxe, G. Labaka, E. Agirre, Translation artifacts in cross-lingual transfer learning, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7674–7684. URL: https://aclanthology.org/2020.emnlp-main.618/. doi:10.18653/v1/2020.emnlp-main.618.

[16] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493/. doi:10.18653/v1/P19-1493.

[17] S. Schuster, S. Gupta, R. Shah, M. Lewis, Cross-lingual transfer learning for multilingual task oriented dialog, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3795–3805. URL: https://aclanthology.org/N19-1380/. doi:10.18653/v1/N19-1380.

[18] J. Khatri, V. Srivastava, L. Vig, Can you translate for me? code-switched machine translation with large language models, in: J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A. A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 83–92. URL: https://aclanthology.org/2023.ijcnlp-short.10/. doi:10.18653/v1/2023.ijcnlp-short.10.

[19] F. Team, U. Abbas, M. S. Ahmad, F. Alam, E. Altinisik, E. Asgari, Y. Boshmaf, S. Boughorbel,

S. Chawla, S. Chowdhury, et al., Fanar: An arabic-centric multimodal generative ai platform, 2025. URL: https://arxiv.org/abs/2501.13944. arXiv:2501.13944.

[20] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, 2024. URL: https://arxiv.org/abs/2403.08295. arXiv:2403.08295.

[21] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[22] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[23] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., Phi-4 technical report, 2024. URL: https://arxiv.org/abs/2412.08905. arXiv:2412.08905.

[24] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[25] N. Cohen-Inger, Y. Elisha, B. Shapira, L. Rokach, S. Cohen, Forget what you know about llms evaluations – llms are like a chameleon, 2025. URL: https://arxiv.org/abs/2502.07445. arXiv:2502.07445.

[26] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 878–891. URL: https://aclanthology.org/2022.acl-long.62/. doi:10.18653/v1/2022.acl-long.62.

[27] Y. Chen, Q. Li, R. Biswas, J. Bjerva, Large language models are easily confused: A quantitative metric, security implications and typological analysis, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Findings of the Association for Computational Linguistics: NAACL 2025, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 3810–3827. URL: https://aclanthology.org/2025.findings-naacl.210/. doi:10.18653/v1/2025.findings-naacl.210.