# No One Fits All: From Fixed Prompting to Learned Routing in Multilingual LLMs

**Wei-Chi Wu**[α] **Sheng-Lun Wei**[α] **Hen-Hsen Huang**[β] **Hsin-Hsi Chen**[αγ]
[α]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan
[β]Institute of Information Science, Academia Sinica, Taiwan
[γ]AI Research Center (AINTU), National Taiwan University, Taiwan
wcwu@csie.ntu.edu.tw, weisl@nlg.csie.ntu.edu.tw,
hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

## Abstract

Translation-based prompting is widely used in multilingual LLMs, yet its effectiveness varies across languages and tasks. We evaluate prompting strategies across ten languages of different resource levels and four benchmarks. Our analysis shows that no single strategy is universally optimal. Translation strongly benefits low-resource languages even when translation quality is imperfect, high-resource languages gain little, and prompt-based self-routing underperforms explicit translation. Motivated by these findings, we formulate prompting strategy selection as a learned decision problem and introduce lightweight classifiers that predict whether native or translation-based prompting is optimal for each instance. The classifiers achieve statistically significant improvements over fixed strategies across four benchmarks and generalize to unseen task formats not observed during training. Further analysis reveals that language resource level, rather than translation quality alone, determines when translation is beneficial.

## 1 Introduction

Translation-based prompting, which translates inputs into English prior to inference, is a widely used strategy for multilingual large language models (LLMs) and often improves performance by leveraging stronger English-centric capabilities (Ghosh et al., 2025). However, recent studies show that this advantage is not universal. Native-language prompting can outperform translation-based approaches on culturally grounded tasks (Tam et al., 2025; Nyandwi et al., 2025) and for models with reduced English bias (Liu et al., 2025). These findings challenge the assumption that translation into English is always beneficial, raising a fundamental question: *when should translation be applied, and when is native-language prompting preferable?*

Prior work has largely focused on improving individual prompting paradigms rather than understanding or selecting between them. Methods such as QAlign (Zhu et al., 2024) and mCoT (Lai and Nissim, 2024) enhance translation-based prompting, while Strategic CoT (Wang et al., 2024) improves native language reasoning. However, these approaches implicitly assume a fixed prompting strategy and do not treat prompting strategy selection as a decision problem conditioned on the language and task pair.

This gap motivates three research questions. (**RQ1**) *Does one prompting strategy fit all languages and tasks?* Through a systematic comparison across diverse languages and tasks, we find that no single strategy consistently dominates. Translation-based prompting benefits low-resource languages but often provides limited or no gains for high-resource languages, while prompt-based self routing yields only marginal improvements and underperforms explicit translation. (**RQ2**) *Can prompting strategy selection be learned?* We formulate strategy selection as a learned decision problem and introduce a lightweight classifier that predicts whether simple native language or translation-based prompting, the two most iconic prompting strategies, is more effective for a given language and task pair. Consequently, the lightweight classifier consistently outperforms isolated strategies across models and task formats. (**RQ3**) *Why does translation primarily benefit low-resource languages?* We show that translation effectiveness is driven more by language resource level than translation quality alone, with the learned selector favoring translation for low-resource languages even when translation quality is imperfect. In summary, our contributions are threefold: **1)** We present a systematic empirical study demonstrating that no single prompting strategy fits all languages and tasks. **2)** We introduce a decision-oriented framework for learned prompting strategy selection. **3)** We provide an analysis uncovering the central role of language resource level in determining when translation-based prompting is beneficial.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 37404–37423
July 2-7, 2026 

| Prompt Method | ZH | ES | DE | HI | BN | ID | KO | SI | SW | YO | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NATIVE | 88.2 | 89.4 | 88.2 | 84.6 | 83.1 | 88.0 | 36.7 | 75.3 | 75.5 | 45.2 | 75.4 |
| TRANSLATE | 87.5 | 89.0 | 88.4 | **86.2** | 85.8 | **88.4** | **86.6** | **82.7** | **81.0** | 64.0 | 84.0 |
| SEL-TRANS | **88.8** | 89.3 | 88.6 | 85.7 | 85.4 | 88.2 | **86.6** | 80.8 | 79.6 | **64.4** | 83.7 |
| SCOT-NATIVE | 85.8 | 86.5 | 84.3 | 83.2 | 71.3 | 82.5 | 77.6 | 71.0 | 65.2 | 46.4 | 75.4 |
| SCOT-TRANS | 88.3 | **89.5** | **88.9** | 85.7 | **87.1** | 88.3 | 86.4 | 81.1 | 80.9 | 63.5 | **84.0** |
| PROMPT-ROUTING | 87.5 | 89.2 | 88.2 | 85.5 | 85.2 | 88.0 | 78.5 | 81.7 | 80.9 | 62.8 | 82.8 |

Table 1: Accuracy (%) of six prompting strategies across ten languages on Global-MMLU using Llama3.3-70B. Languages are grouped by resource level into high (ZH, ES, DE, HI), mid (BN, ID, KO), and low (SI, SW, YO).

## 2 Related Work

**Translation-Based Prompting.** English chain-of-thought reasoning often outperforms native approaches due to English dominance in pretraining (Li et al., 2024; Kowtal et al., 2024). Recent methods improve the issue through question alignment (Zhu et al., 2024), multilingual CoT reasoning (Lai and Nissim, 2024), and instruction tuning with small set (Shaham et al., 2024). Translation effectiveness correlates positively with quality, as low-quality translation can harm performance (Liu et al., 2025).

**Limitations and Alternatives.** Translation fails for culturally grounded tasks (Tam et al., 2025; Nyandwi et al., 2025), models with reduced English bias (Liu et al., 2025), and certain task structures (Huang et al., 2023; Intrator et al., 2024). Alternatives include Strategic CoT (Wang et al., 2024) and Selective Translation (Kowtal et al., 2024; Mondshine et al., 2025; Paul et al., 2025). We learn to select between strategies, revealing that language resource level and response features, not translation quality alone, determine optimality.

## 3 Experimental Setup

**Datasets and Languages.** We primarily evaluate on Global-MMLU (Singh et al., 2025), grouping languages by resource level into high (Chinese/ZH, Spanish/ES, German/DE, Hindi/HI), mid (Bengali/BN, Indonesian/ID, Korean/KO), and low (Sinhala/SI, Swahili/SW, Yoruba/YO). For strategy selection, we use a 10% training split with balanced language coverage and evaluate on the remaining 90%. Generalization is assessed on MMLU-ProX (Xuan et al., 2025) and out-of-domain benchmarks with different task formats: XQuAD (Artetxe et al., 2020), mCSQA (Sakai et al., 2024), and XCOPA (Ponti et al., 2020).

**Prompting Strategies.** We compare zero-shot native and translation-based prompting strategies, including NATIVE, TRANSLATE, SEL-TRANS (Mondshine et al., 2025), Strategic CoT in native and English (Wang et al., 2024), and PROMPT-ROUTING. Prompt templates and details are provided in Appendix A.2.

**Models.** Experiments are conducted using DeepSeek-v3.1 (DeepSeek-AI, 2024), with additional strategy selection experiments on Llama-3.3-70B-Instruct (AI@Meta, 2024). All models are used in zero-shot inference.

**Learned Strategy Selection.** We formulate strategy selection as a binary decision between NATIVE and TRANSLATE. Training labels are assigned when exactly one strategy answers correctly; ambiguous cases are discarded. We train lightweight classifiers (XGBoost (Chen and Guestrin, 2016), MLP (Haykin, 1994)) using features capturing differences between native and translated inputs and responses. Details are in Appendix B.1.

**Features Engineering.** For each instance, we run both NATIVE and TRANSLATE to obtain responses $r_n$ and $r_t$, then extract features capturing their differences across four categories: (1) metadata, (2) question-level, (3) response-level, and (4) alignment. The same language-agnostic pipeline is applied uniformly to all instances. Complete feature definitions appear in Appendix B.2.

## 4 RQ1: Does One Strategy Fit All?

Building on prior findings that question the universality of translation-based prompting, we examine whether any single prompting strategy outperforms others across languages and tasks, as implied by a "one-strategy-fits-all" assumption. Table 1 reveals three key findings. First, **no single strategy dominates**: while SCOT-TRANS achieves the highest average (83.97%), SEL-TRANS wins for 3 languages (ZH, KO, YO) and TRANSLATE for 5 others (HI, ID, KO, SI, SW). Second, **resource level**

| | | *High-Resource* | | | | *Mid-Resource* | | | *Low-Resource* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **ZH** | **ES** | **DE** | **HI** | **BN** | **ID** | **KO** | **SI** | **SW** | **YO** | **Avg** |
| GLOBAL-MMLU | NATIVE | 86.4 | 86.4 | 84.6 | 83.1 | 79.9 | 85.3 | 36.1 | 72.0 | 71.7 | 39.0 | 72.5 |
| | TRANSLATE | 86.0 | 86.8 | 85.5 | 84.4 | 83.0 | 86.0 | 84.8 | 80.0 | 79.2 | 61.6 | 81.7 |
| | CLASSIFIER (Ours) | **86.8** | **87.3** | **86.0** | **84.7** | **83.3** | **86.4** | **84.9** | **80.0** | **79.8** | **63.7** | **82.3** |
| | ORACLE | 90.5 | 90.6 | 91.2 | 89.1 | 88.4 | 90.8 | 89.5 | 86.5 | 85.3 | 72.8 | 88.3 |
| MMLU-PROX | NATIVE | 80.5 | 80.8 | 79.7 | 77.9 | 75.4 | 80.1 | 35.7 | – | 69.3 | 43.4 | 69.2 |
| | TRANSLATE | 80.3 | 81.0 | 80.3 | 79.0 | 78.5 | 80.5 | **79.5** | – | 75.6 | 64.5 | 77.7 |
| | CLASSIFIER (Ours) | **80.8** | **81.3** | **80.5** | **79.2** | **78.7** | **80.7** | **79.5** | – | **76.0** | **64.6** | **77.9** |
| | ORACLE | 85.4 | 85.5 | 85.0 | 84.2 | 83.3 | 85.5 | 82.8 | – | 80.6 | 70.8 | 82.6 |
| XQUAD | NATIVE | 86.6 | 87.2 | 89.2 | 83.6 | – | – | – | – | – | – | 86.7 |
| | TRANSLATE | 87.2 | **88.2** | **90.6** | 82.5 | – | – | – | – | – | – | 87.1 |
| | CLASSIFIER (Ours) | **88.6** | 87.9 | 89.2 | **84.7** | – | – | – | – | – | – | **87.6** |
| | ORACLE | 91.0 | 92.0 | 94.1 | 89.8 | – | – | – | – | – | – | 91.7 |
| MCSQA | NATIVE | 27.6 | – | 38.2 | – | – | – | – | – | – | – | 32.9 |
| | TRANSLATE | 28.2 | – | 38.5 | – | – | – | – | – | – | – | 33.4 |
| | CLASSIFIER (Ours) | **28.4** | – | **39.1** | – | – | – | – | – | – | – | **33.8** |
| | ORACLE | 36.8 | – | 45.7 | – | – | – | – | – | – | – | 39.9 |
| XCOPA | NATIVE | 97.0 | – | – | – | – | 95.8 | – | – | 87.4 | – | 93.4 |
| | TRANSLATE | **97.4** | – | – | – | – | **96.8** | – | – | 91.8 | – | 95.3 |
| | CLASSIFIER (Ours) | **97.4** | – | – | – | – | 96.6 | – | – | **93.0** | – | **95.7** |
| | ORACLE | 99.0 | – | – | – | – | 98.6 | – | – | 97.0 | – | 98.2 |

Table 2: Results on DeepSeek-v3.1 across in-domain ( green ) and out-of-domain ( orange ) benchmarks. Best results are **bolded**; ORACLE marks the upper bound where at least one of NATIVE or TRANSLATE succeeds. Empty cells indicate languages not covered by the respective benchmark dataset, as detailed in Table 8.

**predicts strategy effectiveness**: low-resource languages consistently favor translation (SI/SW/YO: +5.5 to +18.8% over native), while high-resource languages show the opposite trend (<1%). Korean presents an extreme case with a 49.9% gap between strategies, suggesting severe underrepresentation in training. Third, **prompt-based strategy selection fails**: PROMPT-ROUTING (82.8%) underperforms simple TRANSLATE (84.0%) , demonstrating that effective strategy selection requires learning from patterns rather than model self-assessment.

## 5 RQ2: Can We Learn to Select?

### 5.1 Problem Formulation

For each question $q$ in language $\ell$, we generate responses using both NATIVE ($r_n$) and TRANSLATE ($r_t$) strategies. Our goal is to train a binary classifier $f(q, r_n, r_t) \rightarrow \{0, 1\}$ that predicts which strategy yields the correct answer, where 0 selects NATIVE and 1 selects TRANSLATE.

### 5.2 Experiment Results

Table 2 reports the performance of the XGBoost classifier on DeepSeek-v3.1 across all benchmarks. Complete results for both DeepSeek-v3.1 and Llama-3.3-70B are provided in Appendix B.4.

**In-domain Performance.** On Global-MMLU test set, CLASSIFIER achieves 82.3% accuracy, outperforming TRANSLATE (+0.6%) and substantially exceeding NATIVE (+9.8%). The classifier captures consistent performance gains across all languages, especially with improvements most pronounced for YO (+2.1% over best baseline). On MMLU-ProX, gains persist (about +0.2% over best baseline), demonstrating robustness to increased difficulty.

**Out-of-domain Generalization.** Despite training only on multiple-choice questions, the classifier generalizes to different task formats. On XQuAD (extractive QA), it achieves 87.6% (+0.5% over best baseline). On XCOPA (causal reasoning), performance reaches 95.7% (+0.4%). Even on mCSQA's challenging examples, the classifier shows modest gains (33.8% vs 33.4%).

**Statistical Significance.** We assess significance using the Wilcoxon signed-rank test (Wilcoxon, 1945). Across all language-dataset pairs on both models, XGBoost significantly outperforms both baselines ($p < 0.001$), while MLP achieves $p < 0.05$. This demonstrates the statistical robustness of learned strategy selection. Detailed calculation formula and results are presented in Appendix C.

| | XGBoost | MLP |
|---|---|---|
| DeepSeek-v3.1 | 1. Word Overlap (34.38%)<br>2. Metadata (24.52%)<br>3. POS (9.97%) | 1. Word Overlap (35.10%)<br>2. Response Quality (12.62%)<br>3. Question-Level (10.55%) |
| Llama-3.3-70B | 1. Word Overlap (56.91%)<br>2. Question-Level (24.62%)<br>3. Response Quality (17.11%) | 1. Word Overlap (36.78%)<br>2. POS (10.95%)<br>3. Question-Level (10.41%) |

Table 3: Top 3 most important feature groups (with importance scores, %) for XGBoost and MLP classifiers on DeepSeek-v3.1 and Llama-3.3-70B.

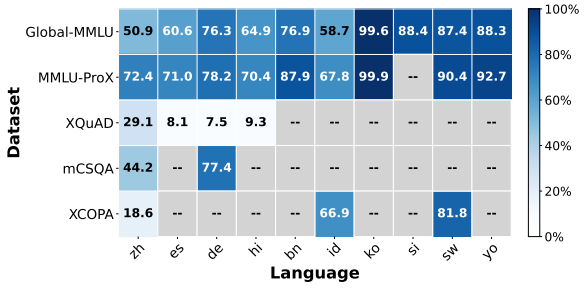


Figure 1: TRANSLATE selection rate (%) of the XGBoost classifier on DeepSeek-v3.1.

### 5.3 Feature Importance Analysis

To further understand what drives the routing decisions, we analyze feature importance of the classifiers. Table 3 shows that word overlap features consistently dominate across settings, suggesting that the classifiers primarily rely on semantic alignment differences between native and translated responses. These features are precisely what prompt-based self-routing cannot access. PROMPT-ROUTING relies on the model's self-assessment, which lacks the ability to quantify response-level differences and features. This explains why PROMPT-ROUTING (82.8%) underperforms simple TRANSLATE (84.0%), while the learned classifier, equipped with these features, consistently outperforms both fixed strategies.

### 5.4 Strategy Selection Analysis

Figure 1 shows the classifier's TRANSLATE selection rate strongly correlates with language resource level: relatively high-resource languages (ZH, ES, DE, HI, ID) exhibit balanced selection (40-70%) varying by task, while relatively low-resource languages (KO, SI, YO) heavily favor TRANSLATE. This contradicts expectations from prior work, which indicates that translating low-resource languages leads to low translation quality (Koehn and Knowles, 2017; Team et al., 2022; Shu et al., 2024) and harm performance (Liu et al., 2025). We therefore investigate this relationship further in RQ3 (§6). Full translation rate heatmaps appear in Appendix D.1.

## 6 RQ3: Why Low-Resource Languages Favor Translation?

We conduct the analysis to explore the relationship among language resource level, translation quality, and learned strategy selection.

**Setup.** We evaluate translation quality using BLEURT (Sellam et al., 2020), chrF (Popović, 2015), and METEOR (Banerjee and Lavie, 2005), comparing model-generated translations against original English questions and options. We partition Global-MMLU and MMLU-ProX examples into quality deciles and measure: (1) accuracy for each method, (2) performance gap (TRANSLATE − NATIVE), and (3) classifier's TRANSLATE selection rate. Details appear in Appendix D.2.

**Results.** Table 4 shows results using chrF on Global-MMLU for DeepSeek-v3.1 and Llama-3.3-70B. Complete results across quality metrics and datasets are provided in Appendix D.2. Three consistent patterns emerge. As translation quality improves: (1) all methods achieve higher accuracy, (2) the TRANSLATE−NATIVE gap narrows, and (3) the classifier's TRANSLATE selection rate correspondingly decreases. Critically, the classifier selects TRANSLATE most aggressively where translation quality is *lowest*, not highest. This inverse correlation demonstrates the classifier learns to effectively exploit translation where native performance is weakest, independent of translation quality itself.

**Discussion.** This pattern reflects the confounding between language resource level, translation quality, and the learned strategy selection by classifiers. As shown in Figure 2, the responses off low-resource languages concentrate in low-quality bins due to limited parallel corpora (Koehn and

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 53.2 | 50.3 | 67.3 | 57.6 | 69.3 | 58.2 | 14.2 | 7.3 | 83.5 | 76.3 |
| 20% | 57.4 | 52.2 | 71.4 | 59.7 | 72.8 | 60.4 | 14.0 | 7.5 | 81.9 | 72.0 |
| 30% | 60.3 | 55.2 | 73.7 | 62.3 | 74.8 | 63.2 | 13.5 | 7.1 | 80.1 | 67.1 |
| 40% | 62.1 | 57.7 | 75.0 | 64.4 | 75.9 | 65.3 | 12.9 | 6.7 | 78.6 | 63.2 |
| 50% | 64.1 | 59.7 | 76.3 | 66.1 | 77.1 | 66.9 | 12.2 | 6.4 | 77.6 | 60.3 |
| 60% | 65.8 | 61.3 | 77.4 | 67.4 | 78.1 | 68.3 | 11.6 | 6.0 | 76.6 | 58.0 |
| 70% | 67.5 | 62.9 | 78.5 | 68.7 | 79.2 | 69.6 | 11.0 | 5.8 | 76.0 | 56.0 |
| 80% | 69.2 | 64.3 | 79.5 | 69.8 | 80.1 | 70.7 | 10.3 | 5.6 | 75.5 | 53.8 |
| 90% | 70.8 | 65.7 | 80.5 | 70.9 | 81.1 | 71.9 | 9.8 | 5.2 | 75.2 | 51.6 |
| 100% | 72.5 | 67.2 | 81.7 | 72.1 | 82.3 | 73.1 | 9.3 | 4.9 | 75.2 | 49.2 |

Table 4: Translation quality analysis on Global-MMLU using chrF scores of the XGBoost classifier. Low quality bins (bottom 30%) show high TRANSLATE selection rates despite lower accuracy. High quality bins (top 40%) show improved accuracy but lower translation rate.

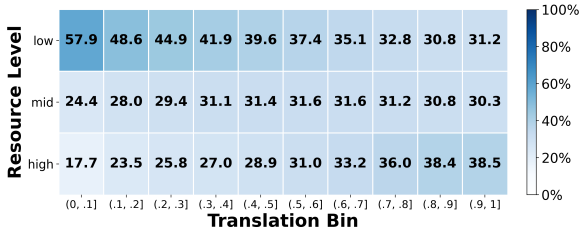


Figure 2: Distribution (%) of responses across translation quality bins by language resource level on Global-MMLU with DeepSeek-v3.1.

Knowles, 2017; Team et al., 2022), but benefit the most from translation (§4). The strategy performance gap narrows as high-resource languages dominate high-quality bins with little strategy differences. Our analysis reveals that **language resource level, not translation quality alone, determines optimal strategy**. Full language resource and quality bins heatmaps appear in Appendix D.2.

## 7 Conclusion

This work investigates prompting strategy selection for multilingual LLMs, showing that translation-based prompting is not universally beneficial and that no single strategy fits all language–task pairs, with low-resource languages favoring translation despite lower translation quality. To address this variability, we introduce lightweight classifiers that predict the optimal strategy for each instance, achieving statistically significant improvements over both native and translation baselines across four benchmarks and generalizing to unseen task formats. Through controlled analysis, we show that language resource level, rather than translation quality, is the primary factor determining when translation is beneficial. These findings reframe multilingual prompting from a fixed-strategy paradigm to a learned decision problem. Future work can build on this through stronger routing models, hybrid prompting strategies, retrieval-based selection methods, and ultimately integrating routing directly into model inference to eliminate dual-generation overhead.

## Limitations

While our classifier demonstrates effectiveness across multiple benchmarks, several limitations warrant consideration. First, our experiments focus on ten languages spanning different resource levels, and the generalizability of our findings to other unseen languages and additional model families, particularly non English-centric models, remains to be validated. Second, although our evaluation already covers a range of task types, it still does not fully represent the diversity of multilingual NLP applications due to the limitations of current existing multilingual datasets. Additionally, the lack of culturally sensitive multilingual datasets makes it difficult to assess whether cultural factors play a role in prompting strategies. Third, the classifier relies on features extracted from model responses, meaning it requires generating both native and translated outputs for each inference, which increases computational costs compared to selecting a single strategy. This overhead may limit practical deployment in resource-constrained settings. We expect that the strategy decision classifier could be further utilized within LLMs to automatically select the responding route without requiring dual inference, potentially through integration as an internal routing mechanism or by training the model to predict optimal strategies based on input features alone.

## Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grant NSTC 114-2221-E-002 -070 -MY3, and by Ministry of Education (MOE), Taiwan, under grant NTU-114L900901.

## References

AI@Meta. 2024. Llama 3 model card.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *Preprint*, arXiv:1907.10902.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. 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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. A survey of multilingual reasoning in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8920–8936, Suzhou, China. Association for Computational Linguistics.

Simon Haykin. 1994. *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 829–844, Mexico City, Mexico. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Nidhi Kowtal, Tejas Deshpande, and Raviraj Joshi. 2024. Chain-of-translation prompting (CoTR): A novel prompting technique for low resource languages. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 645–655, Tokyo, Japan. Tokyo University of Foreign Studies.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *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 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The impact of prompt translation strategies across languages and tasks in multilingual LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico. Association for Computational Linguistics.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual

multimodal LLMs with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24198–24242, Suzhou, China. Association for Computational Linguistics.

Rakesh Paul, Anusha Kamath, Kanishk Singla, Raviraj Joshi, Utkarsh Vaidya, Sanjay Singh Chauhan, and Niranjan Wartikar. 2025. Aligning large language models to low-resource languages through llm-based selective translation: A systematic study. *Preprint*, arXiv:2507.14304.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *Preprint*, arXiv:2411.11295.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2025. Language matters: How do multilingual input and reasoning paths affect large reasoning models? *Preprint*, arXiv:2505.17407.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yu Wang, Shiwan Zhao, Zhihu Wang, Heyuan Huang, Ming Fan, Yubo Zhang, Zhixing Wang, Haijun Wang, and Ting Liu. 2024. Strategic chain-of-thought: Guiding accurate reasoning in LLMs through strategy elicitation. *Computing Research Repository*, arXiv:2409.03271.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

# A Preliminary Details

## A.1 LLM Endpoints

We use the NVIDIA NIM APIs[1] to generate responses for each prompting strategy, using `deepseek-ai/deepseek-v3_1` with thinking mode enabled for DeepSeek-v3.1 and `meta/llama-3_3-70b-instruct` for Llama-3.3-70B.

## A.2 Prompt Strategies

We assess multiple strategies based on the language used for instructions and reasoning steps, using a zero-shot approach. The complete prompting templates are provided in Figure 3, 4, 5, 6, 7, and 8. The prompt templates refer to Liu et al. (2025) (NATIVE, TRANSLATE), Mondshine et al. (2025) (SEL-TRANS), and Wang et al. (2024) (SCOT-NATIVE, SCOT-TRANS).

**Native method (NATIVE)** In NATIVE, we provide the question with both input and Chain-of-Thought instructions in the native language.

**Translate method (TRANSLATE)** In TRANSLATE, we provide the question with input in the native language, then instruct the model to translate the question to English and solve it with English Chain-of-Thought instructions.

**Selective Translate method (SEL-TRANS)** SELECTIVE TRANSLATION (Mondshine et al., 2025) is a method that selectively translates only specific parts of the prompt. We provide the question with input in the native language, then instruct the model using English Chain-of-Thought instructions without first translating the question.

**Native Strategic Chain-of-Thought method (SCOT-NATIVE)** NATIVE STRATEGIC CHAIN-OF-THOUGHT (Wang et al., 2024) is a method that integrates strategic knowledge before generating intermediate reasoning steps. We provide both the input and Strategic Chain-of-Thought instructions in the native language.

**Translate Strategic Chain-of-Thought method (SCOT-TRANS)** In TRANSLATE STRATEGIC CHAIN-OF-THOUGHT, we provide the input in the native language, then instruct the model with English Strategic Chain-of-Thought instructions without first translating the complete question input.

[1] https://build.nvidia.com

**Prompt Routing method (PROMPT-ROUTING)** In PROMPT-ROUTING, we provide the input in the native language, then instruct the model to determine whether to translate the question into English and solve it with native or English Chain-of-Thought instructions.

## A.3 Complete Preliminary Analysis Results

We conduct our preliminary analysis on the Global-MMLU subsets labeled as Culturally Sensitive (CS) and Culturally Agnostic (CA). The main results of the preliminary analysis separated by subsets are presented in Table 5.

# B Classifiers Details

## B.1 Classifier Settings and Details

We employ hyperparameter tuning to optimize the performance of our classifier models during the training session. We use Optuna (Akiba et al., 2019) to perform automated hyperparameter optimization, optimizing overall accuracy (problem-level correctness) as the primary objective. The final hyperparameter values selected for each model configuration are presented in Table 6.

**XGBoost** We tune the number of estimators (100–600), maximum tree depth (3–12), learning rate (0.01–0.3, log scale), subsample ratio (0.6–1.0), column subsample ratio (0.6–1.0), and minimum child weight (1.0–10.0).

**MLP** We tune the hidden layer architecture (selected from predefined configurations), L2 regularization parameter $\alpha$ (1e-5–1e-2, log scale), and initial learning rate (1e-4–1e-2, log scale).

## B.2 Features

Our approach relies on features capturing differences between $r_n$ and $r_t$ across linguistic quality, complexity, and alignment dimensions. Complete feature descriptions and examples appear in Table 7.

**Metadata Features.** Language identifier and subject category provide coarse-grained context about resource availability and domain-specific requirements.

**Question-Level Features.** Punctuation mark count and numeric character count capture structural properties that may interact differently with translation.

*Native Prompting*

[Multiple-choice]
Answer the following multiple choice question. The last line of your response should be exactly: 'Answer $LETTER' where LETTER is one of ABCD. Think step by step before answering.

Question: {question}

Options: {options}

[QA]
Answer the following question based on the given context. Provide a concise and accurate answer. The last line of your response should be exactly: 'Answer: [your answer]'.

Context: {context}

Question: {question}

Figure 3: Native prompting template for LLM response generation. We use Google Translate to translate the instruction into other native languages when prompting.

*Translate Prompting*

[Multiple-choice]
First, translate the following question and options from {language} to English. Then, answer the translated multiple choice question. The last line of your response should be exactly: 'Answer $LETTER' where LETTER is one of ABCD. Think step by step before answering.

Original Question ({language}): {question}

Original Options ({language}): {options}

Please provide your response in the following format:
Translated Question: [your English translation]
Translated Options: [your English translation]
Reasoning: [your step-by-step reasoning]
Answer [LETTER]

[QA]
First, translate the following context and question from {language} to English. Then, answer the translated question based on the translated context. The last line of your response should be exactly: 'Answer: [your answer]'.

Original Context ({language}): {context}

Original Question ({language}): {question}

Please provide your response in the following format:
Translated Context: [your English translation]
Translated Question: [your English translation]
Reasoning: [your step-by-step reasoning]
Answer: [your answer]

Figure 4: Translate prompting template for LLM response generation. All languages use the same English instruction to translate and response.

*Selective-translate Prompting*

[Multiple-choice]
Answer the following multiple choice question. The last line of your response should be exactly: 'Answer $LETTER' where LETTER is one of ABCD. Think step by step before answering.

Question: {question}

Options: {options}

Figure 5: Selective translate prompting template for LLM response generation. All languages use the same English instruction with native inputs.

*Native Strategic CoT Prompting*

[Multiple-choice]
**Role:** You are a strategic reasoning expert skilled in systematic problem-solving.

**Workflow:**
1. First, analyze the problem and develop a strategic approach to solve it.
2. Then, apply your strategy step-by-step to reach the solution.

**Rule:** Ensure each step is logical, clear, and builds upon the previous one.

**Initialization:** Let's begin by understanding the problem and formulating a strategy.

**Task Input:**
Question: {question}

Options: {options}

Please follow the SCoT methodology: first outline your strategic approach, then apply it step-by-step. End your response with exactly: 'Answer $LETTER' where LETTER is one of ABCD.

Figure 6: Native strategic CoT prompting template for LLM response generation. We use Google Translate to translate the instruction into other native languages when prompting.

*Translate Strategic CoT Prompting*

[Multiple-choice]
**Role:** You are a strategic reasoning expert skilled in systematic problem-solving.

**Workflow:**
1. First, analyze the problem and develop a strategic approach to solve it.
2. Then, apply your strategy step-by-step to reach the solution.

**Rule:** Ensure each step is logical, clear, and builds upon the previous one.

**Initialization:** Let's begin by understanding the problem and formulating a strategy.

**Task Input:**
Question: {question}

Options: {options}

Please follow the SCoT methodology: first outline your strategic approach, then apply it step-by-step. End your response with exactly: 'Answer $LETTER' where LETTER is one of ABCD.

Figure 7: Translate strategic CoT prompting template for LLM response generation. All languages use the same English instruction with native inputs.

*Routing Prompting*

[Multiple-choice]
You are a multilingual AI assistant tasked with determining the best approach to answer a multiple-choice question.

Question Language: {language_name}
Question: {question}
Options: {options}

Based on research in multilingual NLP, there are two approaches:
1. NATIVE: Answer directly in {language_name}
2. TRANSLATE: Translate the question to English first, then answer

Please assess your proficiency and confidence:
- How confident are you in understanding and reasoning in {language_name}? (Consider vocabulary, grammar, cultural context)
- Is this a complex question requiring nuanced reasoning, or is it straightforward?
- Would translating to English improve your accuracy?

Respond with EXACTLY ONE of the following on the last line:
ROUTE: NATIVE
or
ROUTE: TRANSLATE

Provide brief reasoning first (1-2 sentences), then your routing decision.

[QA]
You are a multilingual AI assistant tasked with determining the best approach to answer a question based on context.

Question Language: {language_name}
Context: {context}
Question: {question}

Based on research in multilingual NLP, there are two approaches:
1. NATIVE: Answer directly in {language_name} based on the context
2. TRANSLATE: Translate the context and question to English first, then answer

Please assess your proficiency and confidence:
- How confident are you in understanding and reasoning in {language_name}? (Consider vocabulary, grammar, cultural context)
- Is this a complex question requiring nuanced reasoning, or is it straightforward?
- Would translating to English improve your accuracy?

Respond with EXACTLY ONE of the following on the last line:
ROUTE: NATIVE
or
ROUTE: TRANSLATE

Provide brief reasoning first (1-2 sentences), then your routing decision.

Figure 8: Prompt routing template for LLM response generation. All languages use the same English instruction to decide the strategy. After the decision, they use the same prompt as native/translate prompting to get the final response and answer.

| Subset | Prompt Method | ZH | ES | DE | HI | BN | ID | KO | SI | SW | YO | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CS | NATIVE | 85.1 | 86.9 | 85.7 | 80.9 | 78.8 | 85.4 | 37.4 | 69.6 | 72.6 | 40.7 | 72.3 |
| | TRANSLATE | 84.5 | **87.9** | 86.7 | **83.1** | 83.3 | **85.9** | **84.8** | **78.4** | 78.8 | **63.1** | **81.7** |
| | SEL-TRANS | **86.4** | 86.2 | 86.2 | 81.4 | 81.2 | 84.7 | 84.5 | 75.0 | 75.3 | 61.2 | 80.2 |
| | SCOT-NATIVE | 82.7 | 84.7 | 80.7 | 79.8 | 67.2 | 79.0 | 74.9 | 67.4 | 60.9 | 44.2 | 72.2 |
| | SCOT-TRANS | 86.1 | 87.0 | **87.1** | 82.1 | **87.1** | 84.3 | 84.0 | 77.0 | 76.9 | 62.2 | 81.4 |
| | PROMPT-ROUTING | 84.1 | 87.1 | 85.7 | 81.7 | 82.2 | 85.0 | 77.3 | 77.1 | **79.4** | 60.6 | 80.0 |
| CA | NATIVE | 89.4 | 90.4 | 89.2 | 86.1 | 84.8 | 89.0 | 36.5 | 77.5 | 76.6 | 47.0 | 76.7 |
| | TRANSLATE | 88.6 | 89.5 | 89.0 | **87.4** | 86.7 | 89.3 | 87.3 | **84.3** | 81.8 | 64.4 | 84.8 |
| | SEL-TRANS | **89.7** | **90.5** | 89.5 | **87.4** | 87.0 | 89.5 | **87.4** | 83.0 | 81.3 | **65.6** | **85.1** |
| | SCOT-NATIVE | 87.0 | 87.2 | 85.7 | 84.5 | 72.9 | 83.8 | 78.7 | 72.4 | 66.9 | 47.2 | 76.6 |
| | SCOT-TRANS | 89.1 | 90.4 | **89.6** | 87.1 | **87.1** | **89.8** | 87.3 | 82.7 | **82.5** | 64.0 | 85.0 |
| | PROMPT-ROUTING | 88.8 | 90.0 | 89.1 | 86.9 | 86.3 | 89.1 | 78.9 | 83.4 | 81.4 | 63.6 | 83.8 |

Table 5: Accuracy (%) of six prompting strategies on the Culturally Sensitive (CS) and Culturally Agnostic (CA) subsets of Global-MMLU using Llama3.3-70B, across ten languages grouped by resource level.

| Hyperparameter | MLP | | XGBoost | |
|---|---|---|---|---|
| | Deepseek-v3.1 | Llama-3.3-70B | Deepseek-v3.1 | Llama-3.3-70B |
| *MLP Hyperparameters* | | | | |
| Hidden Layer Sizes | $(100, 50)$ | $(100)$ | — | — |
| $\alpha$ (L2 regularization) | $8.94 \times 10^{-5}$ | $4.19 \times 10^{-5}$ | — | — |
| Learning Rate (initial) | $3.27 \times 10^{-3}$ | $5.44 \times 10^{-3}$ | — | — |
| *XGBoost Hyperparameters* | | | | |
| Number of Estimators | — | — | 424 | 101 |
| Max Depth | — | — | 10 | 3 |
| Learning Rate | — | — | $2.87 \times 10^{-2}$ | $1.88 \times 10^{-2}$ |
| Subsample | — | — | 0.951 | 0.700 |
| Column Sample by Tree | — | — | 0.615 | 0.996 |
| Min Child Weight | — | — | 9.51 | 4.71 |

Table 6: Final hyperparameter values for MLP and XGBoost classifiers.

**Response-Level Features.** We compute linguistic quality metrics for both $r_n$ and $r_t$: named entity count using `spaCy` package, rare word ratio (words outside top-10k frequency), grammar fluency score, lexical diversity (type-token ratio), language confidence (probability assigned to detected language) using `langdetect` package, syntactic complexity (average dependency tree depth) using Stanza (Qi et al., 2020) models, and part-of-speech diversity using Stanza models.

**Question-Response Alignment Features.** Word overlap metrics and embedding similarity (cosine similarity of the combinations of question, answer, and response embeddings, using LaBSE (Feng et al., 2022)) measure how well each response addresses the question, potentially revealing translation-induced semantic drift.

### B.3 Training and Evaluation Datasets Details

The complete statistics of the datasets is presented in Table 8.

### B.4 Complete Classifier Results

The complete accuracy results for the XGBoost and MLP classifier for DeepSeek-v3.1 and Llama-3.3-70B are presented in Table 9 and 10.

## C Statistical Significance

The Wilcoxon signed-rank test evaluates whether the median difference between paired observations is zero. For each language-dataset pair $i$, we compute the difference $d_i = s_i^{\text{proposed}} - s_i^{\text{baseline}}$, where $s_i^{\text{proposed}}$ and $s_i^{\text{baseline}}$ are the accuracy scores of the proposed method and baseline, respectively. We then rank the absolute differences $|d_i|$ from smallest to largest, assigning rank $R_i$ to each pair. The test statistic $W$ is computed as:

$$W = \min\left(\sum_{d_i>0} R_i, \sum_{d_i<0} R_i\right)$$

where the first sum is over pairs where the proposed method outperforms the baseline, and the second

sum is over pairs where the baseline performs better. Under the null hypothesis of no difference, $W$ follows a known distribution, from which we derive the $p$-value. The full result is presented in Table 11.

## D Translation Rate Results

### D.1 Complete Translation Rate Heatmaps

The DeepSeek-v3.1 XGBoost classifier's translation rate heatmap is presented in Figure 1 in the main body; the remaining classifiers translation rate heatmaps with different models and classifier's types are presented in Figure 9, 10, and 11.

### D.2 Translation Rate Analysis Experiment

We calculate all translation quality scores by comparing the translated question and options parsed from LLM responses to gold-standard English reference texts. Higher scores indicate translations that more faithfully preserve the meaning and structure of the original English content.

**Results.** The Global-MMLU translation quality analysis results table is in Table 4 in the main body; additional results of translation rate tables across three different quality scores, two datasets, and two models appear in Table 12, 13, 14, 15, and 16. Complete results of distribution of responses across translation quality bins are provided in Figure 12 and 13.

## E Information About Use Of AI Assistants

We use AI assistants only for minimal tasks such as refining text and basic code snippets. All core research, experimental design, and critical examination of the results are performed and verified by humans to ensure the integrity of the process.

| Feature Name | Description | Example |
|---|---|---|
| *Metadata Features* | | |
| language | Language code of the response (e.g., de, zh, es) | "de", "zh" |
| dataset | Name of the dataset (e.g., mmlu_prox, xquad) | "mmlu_prox" |
| subject | Combined subject and category information | "STEM:mathematics" |
| *Question-Level Features* | | |
| question_punct_density | Density of punctuation marks in question text (punctuation count / text length) | Q: "What is 2+2?" (1 punct / 13 chars = 0.08) |
| question_num_density | Density of numeric characters in question text (digit count / text length) | Q: "What is 2+2?" (2 digits / 13 chars = 0.15) |
| *Response-Level Features* | | |
| rare_word_ratio | Proportion of rare words based on corpus frequency (words below median frequency) | "The method uses sophisticated techniques" (rare words: sophisticated, techniques; 2 / 5 = 0.40) |
| named_entity_count | Number of named entities (persons, organizations, locations) detected via spaCy | "Einstein worked at Princeton in Germany" (Einstein, Princeton, Germany = 3) |
| grammar_fluency_score | Overall fluency score accounting for punctuation errors and formatting issues (0.0–1.0, higher is better) | "The answer is correct." (1.0, no errors) vs "The answer is correct.." (0.82, malformed punctuation) |
| grammar_malformed_punct | Count of malformed punctuation patterns (e.g., consecutive marks like "..") | "Is this right??" (1 instance of "??") |
| grammar_missing_final_period | Binary indicator of missing sentence-ending punctuation (1.0 = missing, 0.0 = present) | "The answer is correct" (1.0) vs "The answer is correct." (0.0) |
| lexical_diversity | Type-token ratio measuring vocabulary diversity (unique words / total words) | "The cat sat. The cat ran." (4 unique: the, cat, sat, ran / 6 total = 0.67) |
| language_detection_confidence | Confidence score via langdetect (0.0–1.0, higher = more confident) | "The quick brown fox jumps" (detected as English with 0.95 confidence) |
| language_mismatch | Binary indicator of language mismatch via langdetect (1.0 = mismatch, 0.0 = match) | Expected: English, Detected: Spanish (1.0) |
| syntactic_depth_max | Maximum depth of dependency parse tree (deeper = more complex) | "The cat that the dog chased ran" (deep nesting = depth 6) |
| syntactic_complexity_score | Normalized syntactic complexity (depth / log2(word_count + 1)) | "The book that the student who the teacher praised read" (depth 7, normalized) |
| pos_noun_verb_ratio | Ratio of nouns to verbs (higher = more nominal/informative style) | "The analysis of the data shows results" (4 nouns / 1 verb = 4.0) |
| pos_diversity_unique_tags | Number of unique part-of-speech tags in response | "The cat sat on the mat" (DET, NOUN, VERB, ADP = 4 unique tags) |
| pos_diversity_score | POS diversity score (unique tags / total tags, 0.0–1.0) | "The cat sat" (3 unique / 3 total = 1.0) vs "cat cat cat" (1 unique / 3 total = 0.33) |
| *Alignment Features* | | |
| word_overlap_*_* | Token-level overlap metrics (F1, precision, recall) measuring word overlap between pairs: answer–response, question–answer, and question–response | Response: "The answer is Paris"; Reference: "Paris" (overlap: {paris}, F1 = 0.33) |
| labse_*_*_similarity | Cosine similarity using LaBSE embeddings (cross-lingual semantic similarity) between pairs: answer–response, question–answer, and question–response | "The capital is Paris" (EN) vs "La capital es París" (ES) (similarity = 0.91) |

Table 7: Complete list of features used by the strategy selection classifier, grouped into four categories: metadata, question-level, response-level, and question–response alignment features.

| Dataset | # Examples | Used Languages | Task Type | Notes |
|---|---|---|---|---|
| Global-MMLU | Around 14,000 per language | EN, ZH, ES, DE, HI, BN, ID, KO, SI, SW, YO | Multiple-choice with 4 options | We use 10% of examples to train the classifier (§3). |
| MMLU-ProX | Around 12,000 per language | EN, ZH, ES, DE, HI, BN, ID, KO, SW, YO | Harder multiple-choice format with more than 4 options | |
| XQuAD | 1,190 per language | ZH, ES, DE, HI | Extractive QA | |
| mCSQA | 2,000-6,000 per language | ZH, DE | Commonsense QA | We only extract the examples with `hard` tag. |
| XCOPA | 500 per language | ZH, ID, SW | Causal reasoning | |

Table 8: Statistics of the multilingual benchmark datasets used in our experiments, including the number of examples, covered languages, and task types.

| | | *High-Resource* | | | | *Mid-Resource* | | | *Low-Resource* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **ZH** | **ES** | **DE** | **HI** | **BN** | **ID** | **KO** | **SI** | **SW** | **YO** | **Avg** |
| GLOBAL-MMLU | NATIVE | 86.4 | 86.4 | 84.6 | 83.1 | 79.9 | 85.3 | 36.1 | 72.0 | 71.7 | 39.0 | 72.5 |
| | TRANSLATE | 86.0 | 86.8 | 85.5 | 84.4 | 83.0 | 86.0 | 84.8 | 80.0 | 79.2 | 61.6 | 81.7 |
| | XGBOOST (Ours) | **86.8** | **87.3** | **86.0** | **84.7** | **83.3** | **86.4** | **84.9** | **80.0** | **79.8** | **63.7** | **82.3** |
| | MLP (Ours) | 86.4 | 86.6 | 85.1 | 84.1 | 81.6 | 85.6 | 76.9 | 78.1 | 77.1 | 61.4 | 80.3 |
| | ORACLE | 90.5 | 90.6 | 91.2 | 89.1 | 88.4 | 90.8 | 89.5 | 86.5 | 85.3 | 72.8 | 88.3 |
| MMLU-PROX | NATIVE | 80.5 | 80.8 | 79.7 | 77.9 | 75.4 | 80.1 | 35.7 | – | 69.3 | 43.4 | 69.2 |
| | TRANSLATE | 80.3 | 81.0 | 80.3 | 79.0 | 78.5 | 80.5 | **79.5** | – | 75.6 | 64.5 | 77.7 |
| | XGBOOST (Ours) | **80.8** | 81.3 | 80.5 | 79.2 | **78.7** | **80.7** | **79.5** | – | **76.0** | 64.6 | **77.9** |
| | MLP (Ours) | 80.8 | **81.3** | **80.7** | **79.3** | 78.0 | 79.3 | 78.2 | – | 75.5 | **64.7** | 77.7 |
| | ORACLE | 85.4 | 85.5 | 85.0 | 84.2 | 83.3 | 85.5 | 82.8 | – | 80.6 | 70.8 | 82.6 |
| XQUAD | NATIVE | 86.6 | 87.2 | 89.2 | 83.6 | – | – | – | – | – | – | 86.7 |
| | TRANSLATE | 87.2 | **88.2** | **90.6** | 82.5 | – | – | – | – | – | – | 87.1 |
| | XGBOOST (Ours) | **88.6** | 87.9 | 89.2 | **84.7** | – | – | – | – | – | – | **87.6** |
| | MLP (Ours) | 86.5 | 88.0 | 89.2 | 83.3 | – | – | – | – | – | – | 86.7 |
| | ORACLE | 91.0 | 92.0 | 94.1 | 89.8 | – | – | – | – | – | – | 91.7 |
| MCSQA | NATIVE | 27.6 | – | 38.2 | – | – | – | – | – | – | – | 32.9 |
| | TRANSLATE | 28.2 | – | 38.5 | – | – | – | – | – | – | – | 33.4 |
| | XGBOOST (Ours) | **28.4** | – | **39.1** | – | – | – | – | – | – | – | **33.8** |
| | MLP (Ours) | 28.4 | – | 38.9 | – | – | – | – | – | – | – | 33.7 |
| | ORACLE | 36.8 | – | 45.7 | – | – | – | – | – | – | – | 39.9 |
| XCOPA | NATIVE | 97.0 | – | – | – | – | 95.8 | – | – | 87.4 | – | 93.4 |
| | TRANSLATE | **97.4** | – | – | – | – | 96.8 | – | – | 91.8 | – | 95.3 |
| | XGBOOST (Ours) | **97.4** | – | – | – | – | 96.6 | – | – | **93.0** | – | **95.7** |
| | MLP (Ours) | **97.4** | – | – | – | – | **97.0** | – | – | 89.6 | – | 94.3 |
| | ORACLE | 99.0 | – | – | – | – | 98.6 | – | – | 97.0 | – | 98.2 |

Table 9: Complete results on DeepSeek-v3.1 across in-domain ( green ) and out-of-domain ( orange ) benchmarks. Best results are **bolded**; ORACLE marks the upper bound where at least one of NATIVE or TRANSLATE succeeds. Empty cells indicate languages not covered by the respective benchmark dataset, as detailed in Table 8.

| | | *High-Resource* | | | | *Mid-Resource* | | | *Low-Resource* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Method** | **ZH** | **ES** | **DE** | **HI** | **BN** | **ID** | **KO** | **SI** | **SW** | **YO** | **Avg** |
| GLOBAL-MMLU | NATIVE | 79.5 | 82.4 | **79.7** | 74.6 | 69.8 | 79.6 | 60.5 | 54.4 | 67.1 | 23.9 | 67.2 |
| | TRANSLATE | 78.5 | 80.6 | 76.9 | 76.2 | 72.9 | 78.6 | **76.5** | 67.0 | 69.8 | **43.0** | 72.0 |
| | XGBOOST (Ours) | **80.0** | **82.9** | 79.5 | 76.7 | **73.4** | **80.9** | **76.5** | 68.1 | 69.7 | 43.0 | **73.1** |
| | MLP (Ours) | 79.8 | 82.5 | 79.7 | **76.8** | 73.2 | 80.2 | 69.8 | **69.1** | **70.3** | 40.7 | 72.2 |
| | ORACLE | 85.3 | 87.2 | 85.7 | 83.4 | 81.3 | 85.3 | 82.5 | 76.0 | 79.3 | 55.1 | 80.1 |
| MMLU-PROX | NATIVE | 63.8 | 66.0 | 65.4 | 57.8 | 53.4 | 65.5 | 40.2 | – | 51.6 | 18.8 | 53.6 |
| | TRANSLATE | 57.8 | 59.9 | 58.7 | 54.5 | 52.9 | 59.6 | **56.0** | – | 49.9 | **36.2** | 53.9 |
| | XGBOOST (Ours) | 64.0 | **67.1** | 66.1 | 61.0 | 55.7 | **66.6** | 55.9 | – | 54.2 | 36.0 | **58.5** |
| | MLP (Ours) | **64.3** | 67.0 | **66.2** | **61.0** | **56.5** | **66.6** | 51.3 | – | **56.1** | 35.8 | 58.3 |
| | ORACLE | 71.3 | 72.3 | 71.4 | 67.3 | 64.4 | 72.0 | 63.9 | – | 62.7 | 43.6 | 65.4 |
| XQUAD | NATIVE | 86.1 | 87.4 | 88.4 | 84.5 | – | – | – | – | – | – | 86.6 |
| | TRANSLATE | 83.9 | 86.6 | 88.3 | 80.1 | – | – | – | – | – | – | 84.7 |
| | XGBOOST (Ours) | 85.8 | **87.8** | **89.1** | **85.2** | – | – | – | – | – | – | **87.0** |
| | MLP (Ours) | **86.1** | 87.3 | 88.8 | 84.5 | – | – | – | – | – | – | 86.7 |
| | ORACLE | 90.3 | 91.9 | 92.4 | 88.4 | – | – | – | – | – | – | 90.8 |
| MCSQA | NATIVE | 26.1 | – | 33.8 | – | – | – | – | – | – | – | 30.0 |
| | TRANSLATE | 24.3 | – | 33.5 | – | – | – | – | – | – | – | 28.9 |
| | XGBOOST (Ours) | **26.5** | – | 34.6 | – | – | – | – | – | – | – | **30.6** |
| | MLP (Ours) | 26.1 | – | **35.2** | – | – | – | – | – | – | – | 30.6 |
| | ORACLE | 33.4 | – | 41.8 | – | – | – | – | – | – | – | 36.2 |
| XCOPA | NATIVE | **97.8** | – | – | – | – | 96.6 | – | – | 83.0 | – | 92.5 |
| | TRANSLATE | 97.2 | – | – | – | – | **97.4** | – | – | **89.0** | – | **94.5** |
| | XGBOOST (Ours) | 97.6 | – | – | – | – | 97.0 | – | – | 85.8 | – | 93.5 |
| | MLP (Ours) | **97.8** | – | – | – | – | 97.0 | – | – | 85.8 | – | 93.5 |
| | ORACLE | 99.0 | – | – | – | – | 98.0 | – | – | 94.4 | – | 97.1 |

Table 10: Complete results on Llama3.3-70B across in-domain ( green ) and out-of-domain ( orange ) benchmarks. Best results are **bolded**; ORACLE marks the upper bound where at least one of NATIVE or TRANSLATE succeeds. Empty cells indicate languages not covered by the respective benchmark dataset, as detailed in Table 8.

| **Model** | **Comparison** | **p-value** |
|---|---|---|
| DeepSeek-v3.1 | XGBoost vs Translate | 0.000873 |
| | XGBoost vs Native | 0.000006 |
| | MLP vs Translate | 0.037474 |
| | MLP vs Native | 0.000051 |
| Llama3.3-70B | XGBoost vs Translate | 0.000219 |
| | XGBoost vs Native | 0.000009 |
| | MLP vs Translate | 0.007202 |
| | MLP vs Native | 0.000035 |

Table 11: Wilcoxon signed-rank test p-values combining all datasets.

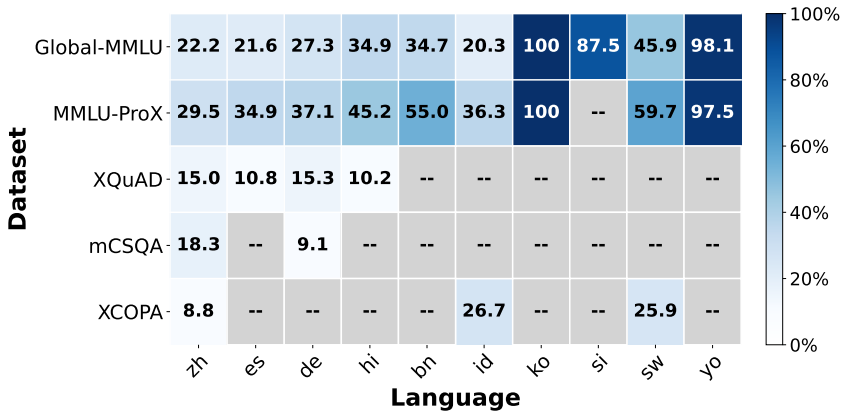


Figure 9: TRANSLATE selection rate (%) of the XGBoost classifier on DeekSeek-v3.1.

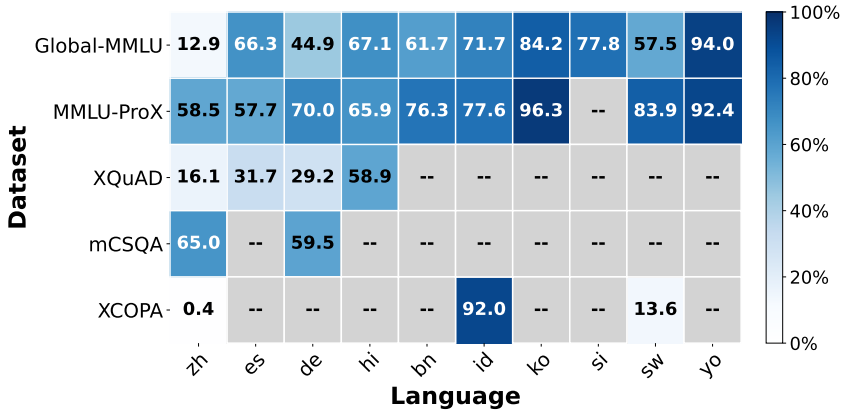


Figure 10: TRANSLATE selection rate (%) of the MLP classifier on DeekSeek-v3.1.

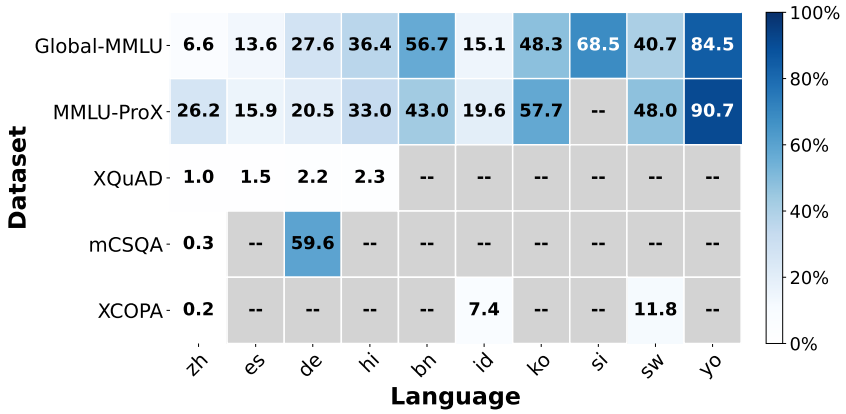


Figure 11: TRANSLATE selection rate (%) of the MLP classifier on Llama-3.3-70B.

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 55.9 | 53.8 | 68.4 | 60.0 | 69.7 | 60.6 | 12.5 | 6.2 | 77.5 | 71.9 |
| 20% | 58.2 | 52.2 | 70.7 | 59.5 | 72.0 | 60.2 | 12.5 | 7.3 | 77.0 | 71.0 |
| 30% | 60.5 | 54.7 | 72.9 | 61.9 | 73.9 | 62.6 | 12.4 | 7.2 | 76.2 | 67.4 |
| 40% | 62.5 | 57.2 | 74.7 | 64.1 | 75.5 | 64.8 | 12.2 | 6.9 | 75.7 | 64.1 |
| 50% | 64.3 | 59.4 | 76.1 | 66.0 | 76.9 | 66.7 | 11.8 | 6.6 | 75.2 | 61.1 |
| 60% | 66.0 | 61.1 | 77.4 | 67.4 | 78.1 | 68.2 | 11.4 | 6.3 | 75.0 | 58.4 |
| 70% | 67.7 | 62.7 | 78.5 | 68.7 | 79.1 | 69.5 | 10.8 | 6.0 | 74.7 | 56.1 |
| 80% | 69.3 | 64.2 | 79.5 | 69.8 | 80.1 | 70.6 | 10.2 | 5.6 | 74.6 | 53.8 |
| 90% | 70.9 | 65.7 | 80.6 | 70.9 | 81.2 | 71.9 | 9.7 | 5.2 | 74.8 | 51.5 |
| 100% | 72.5 | 67.2 | 81.7 | 72.1 | 82.3 | 73.1 | 9.2 | 4.9 | 75.2 | 49.2 |

Table 12: Translation quality analysis on Global-MMLU using BLEURT scores. Low quality bins (bottom 30%) show high TRANSLATE selection rates despite lower accuracy. High quality bins (top 40%) show improved accuracy but lower translation rate.

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 52.5 | 50.5 | 68.2 | 57.9 | 70.0 | 58.6 | 15.7 | 7.4 | 80.0 | 75.3 |
| 20% | 56.6 | 52.1 | 71.3 | 59.8 | 72.5 | 60.4 | 14.7 | 7.7 | 78.8 | 71.9 |
| 30% | 59.5 | 55.1 | 73.2 | 62.2 | 74.2 | 63.0 | 13.7 | 7.1 | 77.5 | 67.3 |
| 40% | 62.0 | 57.4 | 74.8 | 64.1 | 75.7 | 64.9 | 12.8 | 6.7 | 76.6 | 63.3 |
| 50% | 64.1 | 59.5 | 76.2 | 65.8 | 77.0 | 66.7 | 12.1 | 6.3 | 75.9 | 60.3 |
| 60% | 65.9 | 61.3 | 77.2 | 67.3 | 78.0 | 68.2 | 11.3 | 6.0 | 75.5 | 57.7 |
| 70% | 67.6 | 62.8 | 78.3 | 68.6 | 79.0 | 69.5 | 10.7 | 5.8 | 75.1 | 55.4 |
| 80% | 69.2 | 64.4 | 79.4 | 69.9 | 80.0 | 70.8 | 10.2 | 5.5 | 75.0 | 53.3 |
| 90% | 70.8 | 65.8 | 80.6 | 71.0 | 81.1 | 72.0 | 9.8 | 5.2 | 74.9 | 51.4 |
| 100% | 72.5 | 67.2 | 81.7 | 72.1 | 82.3 | 73.1 | 9.2 | 4.9 | 75.2 | 49.2 |

Table 13: Translation quality analysis on Global-MMLU using METEOR scores of the XGBoost classifier of the XGBoost classifier. Low quality bins (bottom 30%) show high TRANSLATE selection rates despite lower accuracy. High quality bins (top 40%) show improved accuracy but lower translation rate.

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 62.0 | 42.7 | 71.9 | 45.8 | 72.4 | 48.4 | 9.9 | 3.1 | 85.8 | 63.7 |
| 20% | 64.1 | 46.7 | 73.3 | 49.0 | 73.7 | 52.4 | 9.2 | 2.3 | 85.2 | 59.3 |
| 30% | 65.3 | 48.1 | 74.5 | 50.0 | 74.8 | 53.9 | 9.2 | 1.9 | 85.0 | 58.2 |
| 40% | 66.0 | 49.1 | 75.3 | 50.9 | 75.6 | 55.0 | 9.3 | 1.8 | 84.5 | 58.1 |
| 50% | 66.1 | 49.6 | 75.5 | 51.2 | 75.7 | 55.6 | 9.4 | 1.6 | 84.0 | 58.1 |
| 60% | 66.4 | 50.3 | 75.8 | 51.6 | 76.0 | 56.0 | 9.4 | 1.3 | 83.6 | 57.8 |
| 70% | 66.6 | 51.0 | 76.0 | 52.1 | 76.3 | 56.5 | 9.4 | 1.1 | 83.1 | 57.5 |
| 80% | 67.2 | 51.7 | 76.4 | 52.7 | 76.6 | 57.1 | 9.2 | 1.0 | 82.5 | 57.0 |
| 90% | 68.0 | 52.6 | 76.9 | 53.2 | 77.2 | 57.7 | 8.9 | 0.6 | 81.8 | 56.1 |
| 100% | 69.2 | 53.6 | 77.7 | 53.9 | 77.9 | 58.5 | 8.5 | 0.3 | 81.2 | 55.0 |

Table 14: Translation quality analysis on MMLU-ProX using BLEURT scores of the XGBoost classifier. Low quality bins (bottom 30%) show high TRANSLATE selection rates despite lower accuracy. High quality bins (top 40%) show improved accuracy but lower translation rate.

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 60.0 | 43.5 | 71.8 | 46.5 | 72.3 | 48.5 | 11.8 | 3.0 | 81.6 | 70.7 |
| 20% | 61.6 | 48.2 | 73.0 | 50.8 | 73.5 | 53.1 | 11.4 | 2.6 | 80.4 | 65.5 |
| 30% | 62.7 | 50.2 | 73.6 | 52.1 | 74.1 | 54.8 | 10.9 | 1.9 | 79.8 | 62.2 |
| 40% | 63.7 | 51.0 | 74.1 | 52.7 | 74.5 | 55.6 | 10.4 | 1.7 | 80.0 | 60.3 |
| 50% | 64.6 | 51.3 | 74.8 | 52.8 | 75.2 | 56.0 | 10.2 | 1.5 | 80.3 | 59.2 |
| 60% | 65.3 | 51.3 | 75.1 | 53.0 | 75.5 | 56.4 | 9.8 | 1.7 | 80.9 | 58.4 |
| 70% | 65.9 | 51.6 | 75.6 | 53.0 | 75.9 | 56.7 | 9.7 | 1.4 | 81.2 | 57.7 |
| 80% | 66.9 | 52.0 | 76.2 | 53.1 | 76.5 | 57.1 | 9.3 | 1.1 | 81.5 | 57.1 |
| 90% | 68.0 | 52.6 | 76.9 | 53.2 | 77.2 | 57.6 | 8.9 | 0.6 | 81.6 | 56.4 |
| 100% | 69.2 | 53.6 | 77.7 | 53.9 | 77.9 | 58.5 | 8.5 | 0.3 | 81.2 | 55.0 |

Table 15: Translation quality analysis on MMLU-ProX using chrF scores of the XGBoost classifier. Low quality bins (bottom 30%) show high TRANSLATE selection rates on Llama-3.3-70B despite lower accuracy. High quality bins (top 40%) show improved accuracy but lower translation rate.

| | *Native* | | *Translate* | | *Classifier* | | *Gap (T-N)* | | *Trans Rate (%)* | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Quality Percentile** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** | **DS** | **Llama** |
| 10% | 61.5 | 47.3 | 73.9 | 50.1 | 74.4 | 52.0 | 12.4 | 2.8 | 79.8 | 68.7 |
| 20% | 62.2 | 50.4 | 73.7 | 52.4 | 74.2 | 54.5 | 11.5 | 2.0 | 78.2 | 63.7 |
| 30% | 62.1 | 50.8 | 72.9 | 52.3 | 73.4 | 54.9 | 10.8 | 1.5 | 78.3 | 61.6 |
| 40% | 62.6 | 50.7 | 72.9 | 52.1 | 73.4 | 55.0 | 10.3 | 1.4 | 78.5 | 60.2 |
| 50% | 63.8 | 50.8 | 73.8 | 52.2 | 74.2 | 55.3 | 10.0 | 1.4 | 79.2 | 59.3 |
| 60% | 64.8 | 51.1 | 74.5 | 52.5 | 74.9 | 55.9 | 9.7 | 1.4 | 80.1 | 58.7 |
| 70% | 66.0 | 51.6 | 75.4 | 52.8 | 75.7 | 56.5 | 9.4 | 1.2 | 80.9 | 57.8 |
| 80% | 67.2 | 52.1 | 76.3 | 53.0 | 76.6 | 57.0 | 9.1 | 0.9 | 81.3 | 56.9 |
| 90% | 68.3 | 52.8 | 77.1 | 53.3 | 77.3 | 57.7 | 8.8 | 0.5 | 81.5 | 56.1 |
| 100% | 69.2 | 53.6 | 77.7 | 53.9 | 77.9 | 58.5 | 8.5 | 0.3 | 81.2 | 55.0 |

Table 16: Translation quality analysis on MMLU-ProX using METEOR scores of the XGBoost classifier. Low quality bins (bottom 30%) show different trend on two models on TRANSLATE selection rates but both have lower accuracy. High quality bins (top 40%) show improved accuracy, while DeepSeek-v3.1 model translation rate increases and Llama-3.3-70B model translation rate drops.

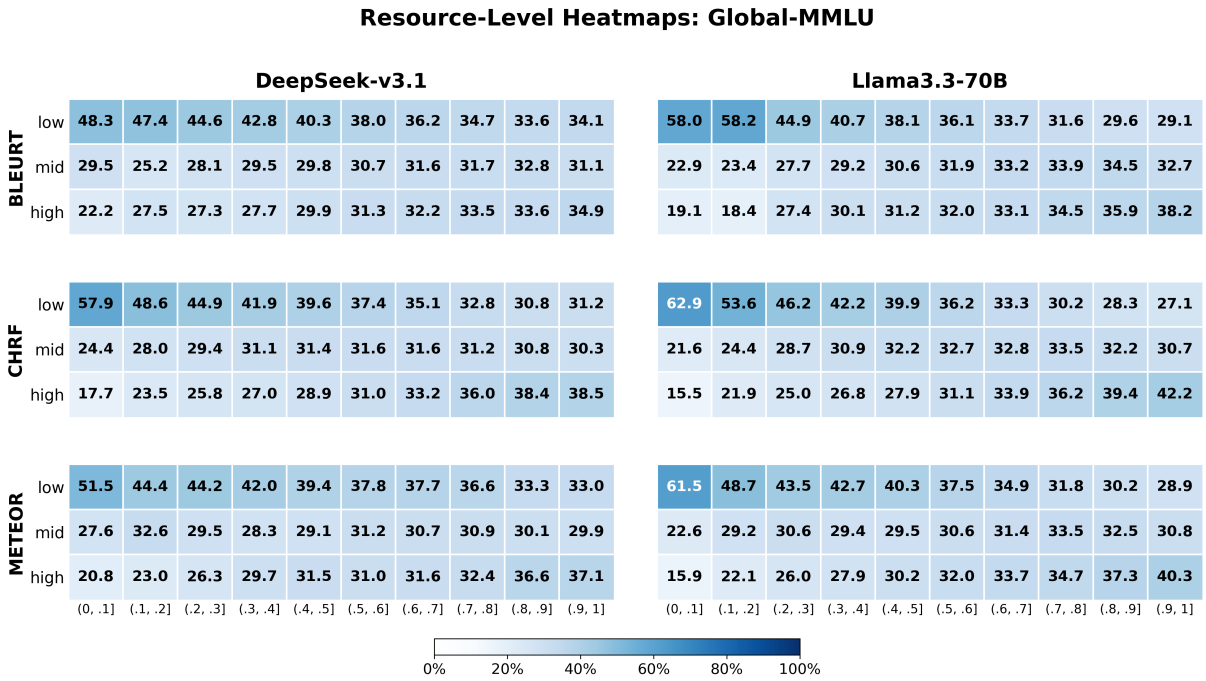


Figure 12: Combined distribution (%) of responses across translation quality bins on Global-MMLU.

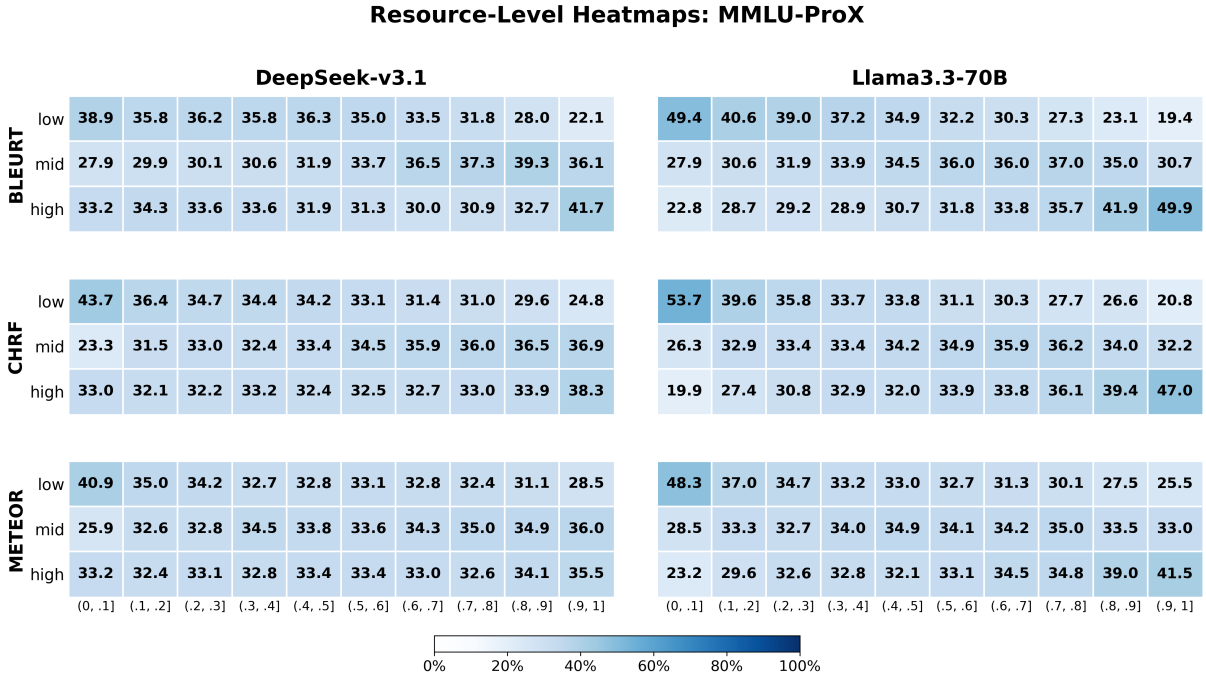


Figure 13: Combined distribution (%) of responses across translation quality bins on MMLU-ProX.