Recovered in Translation: Efficient Pipeline for Automated Translation of Benchmarks and Datasets

Hanna Yukhymenko[1,2], **Anton Alexandrov**[1], **Martin Vechev**[1,2]

[1]INSAIT, Sofia University "St. Kliment Ohridski", [2]ETH Zurich

He broke the ice during the meeting -> He destroyed ice during the meeting

Apple released a new product -> Apple could be Fruit or Technology?

Abstract

The reliability of multilingual Large Language Model (LLM) evaluation is currently compromised by the inconsistent quality of translated benchmarks. Existing resources often suffer from semantic drift and context loss, which can lead to misleading performance metrics. In this work, we present a fully automated framework designed to address these challenges by enabling scalable, high-quality translation of datasets and benchmarks. We demonstrate that adapting test-time compute scaling strategies, specifically Universal Self-Improvement (USI) and our proposed multi-round ranking method, T-RANK, allows for significantly higher quality outputs compared to traditional pipelines. Our framework ensures that benchmarks preserve their original task structure and linguistic nuances during localization. We apply this approach to translate popular benchmarks and datasets into eight Eastern and Southern European languages (Ukrainian, Bulgarian, Slovak, Romanian, Lithuanian, Estonian, Turkish, Greek). Evaluations using both reference-based metrics and LLM-as-a-judge show that our translations surpass existing resources, resulting in more accurate downstream model assessment. We release both the framework and benchmarks to facilitate robust and reproducible multilingual AI development. [1] [2]

1 Introduction

Large Language Models have demonstrated rapid progress in machine translation, outperforming classical tools such as Google Translate and DeepL (Deutsch et al., 2025). This advancement benefits multilingual model development by addressing data and benchmark scarcity in non-English settings. Recent work shows that sampling multiple translation candidates at higher temperatures enhances translation quality, with test-time scaling methods like Best-of-N (Stiennon et al., 2020) and Fusion-of-N (Khairi et al., 2025a) demonstrating improved performance across multilingual domains.

Key challenge: multilingual benchmark quality. A significant gap in recent research is the limited exploration of benchmark translation quality. Multilingual benchmarks are critical for assessing the quality of given model, yet receive less attention than data collection. Existing benchmarks remain imperfect due to reliance on older LLMs or oversimplified methods – most were translated using classical MT tools lacking instruction prompting capabilities, or via older models like GPT-4 whose multilingual capabilities lag behind current frontier models (Lai et al., 2023).

We investigate available benchmarks and identify common translation flaws stemming from translating questions and answers separately, leading to contextual and grammatical mismatches. We focus on Eastern and Southern European languages because: (1) they exhibit complex grammatical features (extensive case systems, grammatical gender, aspect-based verbs) that are sensitive to contextual misalignment; (2) they represent mid-resource languages where frontier models still lack adequate localization; (3) unlike lower-resource languages, translated benchmarks exist for comparison. We translate MMLU, Hellaswag, ARC, and Winogrande into Ukrainian, Romanian, Slovak, Lithuanian, Bulgarian, Turkish, Greek and Estonian.

In this paper, we address three key research questions:

- **RQ1**: Are current translated multilingual benchmarks robust and reliable?
- **RQ2**: To what extent do test-time compute scaling methods improve machine translation quality?
- **RQ3**: How can we translate benchmarks and datasets into other languages while efficiently

Instead of immediately accepting the first answer, give the model more computation at inference time so it can reason, generate alternatives, check itself, or refine its answer

[1]Code: insait-institute/ritranslation

[2]Benchmarks: insait-institute/multilingual-benchmarks

Findings of the Association for Computational Linguistics: ACL 2026, pages 41590–41609
July 2-7, 2026 ©2026 Association for Computational Linguistics

integrating their unique linguistic features into the evaluated tasks?

Our work: To address these questions, we present a novel automated translation framework supporting four methods across various model types, including open-weight models. The framework facilitates machine translation of datasets and benchmarks with minimal manual supervision and maximum configurability, offering multiple methods to balance costs and time investment based on language performance in selected MT models. Our framework incorporates: (1) classic one-prompt translation with optional self-correction (SC) as a lightweight solution; (2) Best-of-N with LLM-as-a-judge scoring to provide reward signals; (3) Universal Self-Improvement (USI) with multi-prompt sampling and unification of translation candidates into refined outputs; (4) our newly proposed Translation Ranking (T-RANK), which employs multi-prompt candidate sampling and multi-round competitive ranking to enhance error detection and achieve superior translation quality.

We evaluate these test-time scaling methods across several mid-resource languages, translate popular benchmarks into these languages, and compare our results with existing translations. Our findings demonstrate that the integrated methods produce higher-quality translations with more accurate evaluation results and improved text quality. Moreover, these methods yield performance improvements on established machine translation benchmarks such as WMT24++ and FLORES.

Main contributions: We summarize our key contributions as follows:

- A comprehensive analysis of multilingual benchmark quality, identifying the root causes of current translation deficiencies.
- A novel, fully automated framework that enables efficient translation of datasets and benchmarks with maximum flexibility. The pipeline is fully configurable, allowing practitioners to optimize the balance between translation quality and cost in multilingual development.
- We release several benchmarks translated into Eastern and Southern European languages, including Ukrainian, Romanian, Slovak, Lithuanian, Bulgarian, Turkish, Greek and Estonian. We find, that translation quality influences benchmark evaluation results significantly, and our proposed methods yield more reliable and accurate multilingual evaluation.

These findings enable more efficient multilingual model development by improving both data and benchmark acquisition.

2 Background and Related Work

2.1 Foundation for LLM-Based Efficient Translation

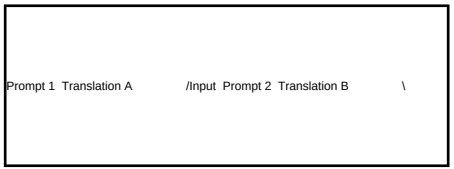
The rapid expansion of multilingual applications for large language models (LLMs) necessitates scalable and high-quality translation frameworks. Recent advancements propose innovative methodologies to enhance LLM-based translation through adaptive prompting and self-refinement strategies.

The Adaptive Few-shot Prompting (AFSP) framework addresses prompt sensitivity in machine translation by dynamically selecting suitable translation demonstrations. AFSP retrieves semantically similar examples from parallel corpora based on LLM-generated embeddings, then generates and reranks multiple translation candidates to select the most contextually appropriate translation (Tang et al., 2025).

The TEaR (Translate, Estimate, and Refine) framework introduces systematic self-refinement for LLM-based translation. TEaR operates by translating the source text, estimating translation quality, and refining based on identified errors. This iterative process enables LLMs to self-correct and improve translation quality, with cross-model experiments demonstrating that general-purpose LLMs can perform both translation and evaluation simultaneously (Feng et al., 2024).

In addition to these frameworks, several test-time compute scaling methods show potential for enhancing LLM translation capabilities. Although originally designed for general reasoning tasks such as mathematics and coding, these methods demonstrate promising applications in translation:

- **Best-of-N Sampling**: This method generates multiple translation outputs and selects the best one based on predefined criteria. It leverages the diversity in LLM outputs due to temperature variations, increasing the likelihood of obtaining high-quality translations (Stiennon et al., 2020).
- **Universal Self-Consistency (USC)**: USC extends the concept of self-consistency by en-



abling LLMs to select the most consistent answer among multiple candidates without relying on answer extraction processes. This approach is particularly effective for open-ended generation tasks, improving performance in mathematical reasoning, code generation, and summarization (Chen et al., 2023).

- **Fusion-of-N**: Rather than selecting the single best answer as in Best-of-N, Fusion-of-N synthesizes the most informative elements from multiple candidates into a single final output. The method uses an LLM judge to aggregate diverse strengths from the candidate pool. Fusion-of-N outperforms Best-of-N and shows promising gains on multilingual tasks, including machine translation (Khairi et al., 2025a).

Findings from Khairi et al. (2025b) also confirm, that sampling multiple candidates with higher temperature with self-improvement and refined selection process leads to significant improvements in many multilingual domains, including machine translation.

Results from the WMT24++ benchmark (Deutsch et al., 2025) further corroborate the efficacy of these methodologies, demonstrating that state-of-the-art LLMs outperform traditional machine translation tools across various language pairs. The research suggests that integrating test-time compute strategies can significantly enhance translation quality, particularly for mid- or low-resource languages. A clear trend emerges whereby newer and larger LLMs consistently outperform existing tools and earlier model iterations, indicating the potential for continued improvements in translation quality as more advanced models are released.

Takeaway 1: Large language models outperform traditional machine translation systems like Google Translate and DeepL through targeted prompts that address domain-specific and language-specific translation challenges.

Collectively, these methodologies underscore the importance of adaptive prompting, self-refinement, and test-time scaling strategies in enhancing LLM-based translation systems. By integrating these approaches, it is possible to develop more robust, efficient, and high-quality translation frameworks capable of addressing the challenges posed by mid- and low-resource languages while minimizing the need for manual human intervention.

2.2 Shortcomings of Existing Multilingual Resources

The rapid development of large language models (LLMs) has highlighted the need for robust multilingual benchmarks. However, translating existing benchmarks often introduces issues that compromise accurate assessment of LLM capabilities across languages.

A prominent example is the MuBench benchmark dataset introduced by Han et al. (2025), comprising widely used benchmarks (Hellaswag (Zellers et al., 2019), ARC (Clark et al., 2018), Winogrande (Sakaguchi et al., 2019), MMLU (Hendrycks et al., 2021), and others) translated into 61 languages with 3.9M samples. While this initiative represents an important step in multilingual evaluation, the translation process relies on an automated pipeline with quality control measures including semantic consistency evaluation, translation purity assessment, and cultural sensitivity checks. However, despite human expert evaluation of 30k samples across 15 languages combined with LLM-as-a-Judge validation, the predominantly automated approach raises concerns regarding translation quality for the remaining samples, as machine translation methods may fail to capture nuanced linguistic and cultural contexts in all 61 languages. Further variations in quality assurance across languages could introduce semantic inaccuracies that skew evaluation results.

Beyond general quality concerns, certain benchmarks present inherent translation challenges. For example, Winogrande, MMLU and Hellaswag may contain answer options with gender-specific adjective endings for sentence completion tasks. In languages with gender-specific adjectives, translating these options can inadvertently reveal correct answers or mislead towards picking an incorrect option, allowing models to succeed through language proficiency rather than reasoning capability. Such linguistic features can compromise evaluation integrity through unintended answer leakage. Therefore, there is a critical need for a flexible framework which allows to adapt translation methods and prompts to address specific issues for different benchmark and language combinations, while MuBench universal approach lacks such flexibility and does not explicitly address these challenges.

The Global-MMLU project (Singh et al., 2024)

represents a substantial effort to advance multilingual evaluation by translating the MMLU benchmark into 42 languages. The process combines machine translation using Google Translate with crowdsourced human verification; however, only approximately 20% of machine-translated texts underwent manual correction. Moreover, questions and answers were translated separately, resulting in observable grammatical inconsistencies in translated entries for languages such as Ukrainian. This context-agnostic translation approach can produce inconsistencies or grammatical errors that affect benchmark coherence and accuracy, particularly for lower-resource languages. Additionally, relying on tools like Google Translate without thorough human validation may overlook language-specific nuances essential for accurate evaluation.

The Okapi framework (Lai et al., 2023) introduces a novel approach to multilingual instruction tuning by leveraging Reinforcement Learning from Human Feedback (RLHF). Unlike traditional methods relying solely on supervised fine-tuning (SFT), Okapi combines SFT with RLHF to align model outputs more closely with human preferences across diverse languages. Moreover, some of the popular benchmarks like MMLU, Hellaswag and ARC were also translated to 26 languages to enable multilingual evaluation. However, Okapi's translation process utilized the ChatGPT model series and the translation could be further improved by employing test-time compute scaling methods to enhance translation quality. Additionally, the framework does not explicitly address language-specific grammatical features that may impact benchmark coherence and accuracy. We present the examples of such translation flaws mentioned before in Appendix A.1.

Takeaway 2: Translating questions and answer options within the same prompt context is essential for sentence completion tasks, as it preserves semantic relationships and prevents contextual mistakes during evaluation.

Recent studies, including WMT24++ (Deutsch et al., 2025), demonstrate that state-of-the-art LLMs outperform traditional machine translation tools such as DeepL or Google Translate across all evaluated language pairs. This finding suggests that reliance on older translation tools may be insufficient for ensuring high-quality translations in multilingual datasets and benchmarks. In addition, practical limitations of popular frameworks (such as low API rate limits or inability to use advanced prompting techniques) can hinder large-scale translation efforts, making it more difficult to produce comprehensive and accurate datasets.

In summary, while initiatives such as MuBench, Global-MMLU, and Okapi represent important progress toward multilingual LLM evaluation, the methods employed in translating benchmarks exhibit limitations that hinder improvement. Dependence on automatic translation without comprehensive human verification, challenges arising from language-specific grammatical structures, and the limited capabilities of translation tools collectively highlight the need for more sophisticated approaches that respect linguistic features. Further, current frontier language models are already outperfoming classical tools, which were used to translate some of the widely used multilingual benchmarks. Future research should advance automated methodologies that reduce reliance on manual curation while simultaneously improving both the quality and accessibility of multilingual evaluation resources.

3 An Efficient Automated Translation Framework

In this section we present a novel automated translation framework, which utilizes Large Language Models with features adapted for both dataset and benchmark (QA/test) formats. Moreover, the framework also allows flexibility in methods and their inner configurable parameters to optimize cost- and time-effectiveness.

3.1 Motivation

As highlighted in previous sections, we observed the lack of research and solutions towards scaled automated translation adapted for custom benchmark formats without loss of translation quality. For this framework, we aim to tackle the following key problems in LLM-assisted data translation:

- **Support for Diverse Data Formats**: Many datasets have complex, nested structures that complicate processing. For LLM training, flat string fields without hierarchical nesting ensure easier ingestion and more efficient workflows.
- **Preservation of Benchmark QA Structure**: Benchmark evaluation reliability depends on

maintaining coherence between questions and answer choices. Preserving these relationships during translation ensures benchmarks remain faithful to the original task structure.

- **Adapting to Varying Language Resource Availability**: High-resource languages (German, Spanish, Hindi) achieve strong translation quality with simple zero-shot prompting, while mid- and low-resource languages require sophisticated, language-specific pipelines. Users need flexibility to configure translation approaches based on each language's characteristics.

- **Addressing Language-Specific Phenomena**: EEU languages possess unique grammatical features requiring careful handling. Language-specific prompt engineering with few-shot examples and LLM-powered verification stages can improve translation fidelity by identifying and correcting language-specific errors.

3.2 Methodology

In our framework, we propose two configuration modes – Dataset and Benchmark. They use different prompts and data format handling, since dataset is more straightforward, but benchmark has a complex connection between question and answers, which needs to be preserved. Finally, we propose four translation methods:

- **SC (Self-Check)**: 0-shot simple translation with optional additional check from another judge LLM.

- **Best-of-N sampling**: sampling N translation candidates and prompting the model to score them, then picking the one with the highest score.

- **USI (Universal Self-Improvement)**: sampling N translations, asking the model to combine them into the best one according to predefined evaluation criteria.

- **T-RANK (Translation Ranking)**: sampling N translations, asking the model to rank them according to their quality and coherence, then correcting the best candidate.

We now examine the translation methods employed in our framework, highlighting their advantages and optimal use cases.

Default translation with optional self-check (SC). This method involves a simple 0-shot prompting of the LLM to translate a text. The user has an option to include a self-check stage: after translation, the model (in a new chat with no history) is prompted to evaluate and correct the result with respect to the original text content. This method is suitable for large text translation using high-resource languages, as it is less costly to perform, due to the sufficient translation capabilities in high-resource languages; however, since the model is prompted to look out for potential errors it might hallucinate them. Additionally, there exists an option to include a few-shot prompt to provide an example of which language-specific points the model should consider during translation.

Best-of-N sampling (BoN). Drawing from test-time compute scaling methods, we implement Best-of-N sampling without a reward model to maintain a training-free framework. We sample N translations at higher temperature (0.7) for diversity, then prompt the LLM to score candidates 1-10 based on specified criteria (Appendix A.5), selecting the highest-scoring translation. While cost-effective and language-agnostic, this method yields lower quality than T-RANK and USI, as LLMs exhibit limitations in numerical scoring and positional bias, favoring earlier candidates despite identifying obvious errors.

Universal Self-Improvement (USI). Building on Universal Self-Consistency and Fusion-of-N, this method operates on the principle that the most consistent translation is not necessarily the best. While Fusion-of-N improves translation quality substantially, the absence of precise translation-specific metrics limits the model's ability to efficiently identify each candidate's strengths for optimal merging, potentially resulting in fusion outputs that incorporate errors. Therefore, we adapt this approach to cultivate self-improvement specifically for machine translation as shown in Appendix A.1 Figure 4. We sample N candidate translations using higher temperature (0.7 recommended), then present them to an evaluator LLM with instructions to combine the candidates into the best version according to specified criteria (shown in Appendix A.5 Table 26). During a combination phase, we ask the model to use only the best features from all seen candidates. This method proves cost-efficient and time-efficient, requiring only $N + 1$ model calls per entry, while successfully addressing and correcting language-specific features. We

illustrate the Universal Self-Improvement workflow in more detail in Figure 1.

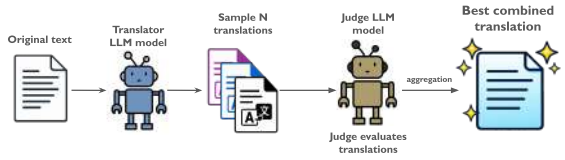


Figure 1: Universal Self-Improvement (USI) method workflow outlook

Translation Ranking (T-RANK). Building upon prior work in translation ranking, we propose an adaptive approach that refines this concept. This method begins by sampling N diverse translations with high temperature. A judge model then evaluates these candidates by ranking them according to predetermined quality criteria, checking general quality, domain consistency, preservation of original question idea and other (shown in Appendix A.5 Table 26), with the top-ranked translation selected as the final output, optionally with refinements. Using this method, we observe more sophisticated reasoning from non-reasoning models, as they successfully identify subtle errors that other methods fail to correct. Moreover, evaluation through comparative ranking rather than numerical scoring appears to facilitate more detailed and rigorous assessment of translation quality.

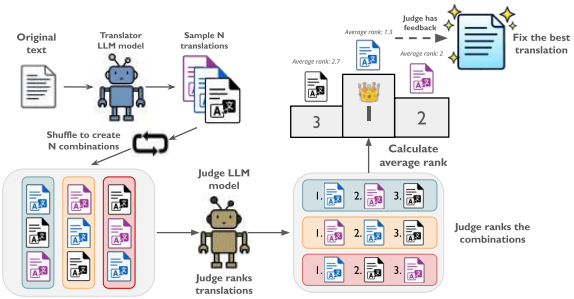


Figure 2: Translation Ranking (T-RANK) method workflow outlook

Previous research on LLM-based evaluation has identified several risks and biases inherent to such methodologies. One notable issue is positional bias: LLMs tend to assign higher scores to candidates presented earlier in the sequence (as shown in Appendix A.1 Table 5). Also, LLM judges often focus their evaluation on the first candidate, identifying specific flaws, and subsequently bias their assessment of following candidates by primarily searching for similar errors, potentially overlooking novel issues in later candidates.

To reduce these prejudices, we introduce a multiple-round sequential ranking strategy, in which the candidate translations are systematically moved across different positional orders. In particular, each translation candidate is presented once in every possible position across all rounds. This process can be visualized as the construction of a sequential grid, where with each new round, candidates are moved one position further till the end (as visualized in Figure 2). After N rounds for N candidates, each candidate has been presented in each position exactly once. Lastly, we show all the candidates to the judge model again and ask to correct and refine the selected translation candidate if needed – this improves model's ability to notice potential translation flaws by looking at strong and weak sides of other variants. We noticed an improved behavior and reasoning when using this method and in particular, presenting *all* candidates to the judge model each time when doing the final correction, as shown in Appendix A.1 Figure 3. This approach demonstrates a promising improvement in reliability of the evaluation by reducing positional biases while controlling evaluation costs and computational efficiency with a total of $2N+1$ model calls. Optionally, once can combine best candidate correction during ranking, however, we found evaluator model's reasoning to be more attentive in a separated setup. Compared to output of USI method shown in Appendix A.1 Figure 3, we can observe that USI method performs fusion of methods, while T-RANK finds more inconsistencies across all candidates and tries to remove them. Since USI fusion step is done only once, the judge model tends to focus on translation flaws seen in first candidates and then tries to check for those specific errors in other candidates. With the multi-round ranking the model is able to see all possible pitfalls from presented options, improving the quality of selection and final improvement even further, while mitigating positional bias at the same time.

We provide example prompts for all proposed methods in Appendix A.5.

Takeaway 3: While Best-of-N and USI sampling improve translation quality, the T-RANK method's competitive ranking approach proves more efficient at identifying subtle translation errors.

3.3 Benchmark Translation for Eastern European Languages

We have translated several popular benchmarks into Eastern European languages using our proposed framework. The selected benchmarks include MMLU, Hellaswag, ARC, and Winogrande, which are widely used for evaluating LLM capabilities across various tasks. We focused on Eastern and Southern European languages such as Ukrainian, Romanian, Slovak, Lithuanian, Greek, Bulgarian, Turkish and Estonian due to their complex grammatical structures and mid-resource status. These languages present unique challenges for machine translation, making them ideal candidates for testing the effectiveness of our framework. MMLU was translated using the GPT-4o-mini-2024-07-18 model checkpoint from OpenAI and all mentioned benchmarks were translated to Ukrainian using the same model. Remaining benchmark translations to other languages were performed using the Gemini-2.0-Flash model from Google. We provide a detailed breakdown of used translation models and methods in Appendix A.2 Table 10. In the next section and in Appendix A.3, we present a comprehensive evaluation of the translation quality achieved through our framework, comparing it with existing translations and assessing its impact on benchmark performance.

4 Evaluation Results

4.1 Machine Translation Benchmarks

We utilize two datasets to evaluate the translation quality of our proposed methods: FLORES and WMT24++.

FLORES. The FLORES benchmark evaluates machine translation systems across multilingual scenarios. FLORES-101 provides 3,001 (1,012 devtest split) sentences from English Wikipedia professionally translated into 101 languages, enabling evaluation of many-to-many translation systems, particularly for low-resource language pairs (Guzmán et al., 2019; Goyal et al., 2022).

WMT24++. The WMT24++ benchmark covers 55 languages and dialects with human-written references and post-edits across four domains: literary, news, social, and speech. This diversity enables comprehensive evaluation of translation systems across high-resource and low-resource languages (Deutsch et al., 2025).

Both FLORES and WMT24++ serve as standard benchmarks in the machine translation community due to their broad language coverage and high-quality human translations, enabling meaningful comparisons across models and approaches.

We evaluate our proposed methods on English-Ukrainian translation using the COMET (Crosslingual Optimized Metric for Evaluation of Translation) metric. COMET leverages multilingual pretrained models to assess translations by comparing source text, hypothesis, and reference, demonstrating higher correlation with human judgments than traditional metrics like BLEU or chrF++ (Rei et al., 2020). We report COMET system-level scores (aggregated across all translations) in Table 1, using the Unbabel/XCOMET-XL model for reference-based quality estimation (Guerreiro et al., 2023). For mentioned tasks we have used GPT-4o-mini-2024-07-18 model for translation.

Method	WMT24++	FLORES
Baseline	0.827	0.937
SC (with check)	0.821	0.937
Best-of-N (n=5)	0.843	0.943
USI (n=5)	0.843	**0.945**
T-RANK (p=5)	**0.845**	0.940

Table 1: COMET (reference-based) translation quality scores for WMT24++ and FLORES EN→UK pair (GPT-4o-mini). n denotes number of sampled candidates, p denotes number of different translation prompts (for $p > 1$ we use $n = 1$).

Our results indicate that USI and T-RANK demonstrate clear advantages over other methods; however, it remains unclear whether T-RANK consistently outperforms USI. This raises the question of whether a correlation exists between translation cost or effort and quality. We must account for the fact that COMET evaluations are not entirely reliable, even when reference translations are provided. Notably, both datasets used for evaluation (WMT24++ and FLORES) primarily contain short texts that do not adequately test complex grammatical structures, particularly for Ukrainian. Moreover, the choice of a correct translation often depends on personal preferences and stylistic conventions; a single source text may have multiple valid translation candidates, all grammatically correct. Therefore, reference translations in commonly used machine translation benchmarks cannot be considered absolute gold standards, and achieving the highest COMET scores does not guarantee that a method is error-free. This motivates the need for

additional comparative tests under different evaluation paradigms, such as Quality Estimation (QE) or reference-free machine translation evaluation. As discussed in appendix A.4, USI is more suitable for short and simple dataset translation, whereas T-RANK shows better performance when translating benchmarks, especially when they have complex question structure, which might get "lost in translation" for languages explored in this paper.

For reference-free QE, we can directly use COMET to score translations without requiring an ideal reference translation. We compared USI and T-RANK methods using the QE setup on FLORES and MMLU datasets, which contain longer and more complex texts. In this work, we use Unbabel/wmt23-cometkiwi-da-xl model for reference-free QE COMET evaluation Rei et al. (2023) and the detailed evaluations are provided in appendix A.4.

We should also note, that using COMET models for machine translation quality evaluation has certain limitations, including bias towards specific domains in training data, technical setup impacting reproducibility, and potential discrepancies with human judgment (Zouhar et al., 2024). While COMET provides a useful automated metric, it is not infallible and should be complemented with human evaluation or using LLM-as-a-judge for comprehensive assessment.

4.2 Multilingual Benchmark Translation Quality

Machine translation quality estimation models struggle to provide consistent and justified feedback on bigger sized texts. Moreover, they are not to able to evaluate the preservation of questions and answer content for the benchmark format, Therefore, we also use LLM-as-a-judge to compare MMLU translations between standard industry-recognized source (Global-MMLU) and our proposed translation methods (T-RANK/USI). MMLU was translated using GPT family of models, thus, we select a model from another family, Gemini-2.5-Flash, as a judge, using which we showcase a clear advantage in translation quality using our proposed methods across Ukrainian, Romanian, and Lithuanian, as presented in Table 2. Notably both methods outperform Global-MMLU alternative in quality of translations, confirming the effectiveness of our framework for benchmark translation.

We also compare evaluation results of mid-sized models like Gemma 3 4/12B, Qwen 3 8B and Llama 3.1 8B on our improved translations versus existing benchmark translations. As shown in Table 3, we observe higher scores on our translations, providing additional evidence of enhanced translation quality. Below, we provide average difference between evaluating on our translations and existing ones (positive difference means results are higher on our translations); more detailed results are presented in Appendix A.3.

Language	Method	Wins	Draws	Losses
UK	T-RANK	**8750**	3276	2016
RO	T-RANK	**8376**	3020	2646
LT	USI	**7382**	3181	3478

Table 2: LLM-as-a-judge comparison between Global-MMLU and our translations (T-RANK/USI) using Gemini-2.5-Flash as judge. Our translations win significantly more comparisons across all evaluated languages.

Benchmark	Languages	AVG Δ
Winogrande	UK, RO, SK, BG, LT, TR, EL, ET	+3.42%
ARC-Challenge	UK, RO, SK, BG	+2.35%
Hellaswag	UK, RO, SK, BG	+1.63%
MMLU	UK, RO, SK, BG, LT, TR, EL	+0.94%

Table 3: Average improvement per benchmark across all languages

We can observe gains across all studied benchmarks, with the most significant improvement in Winogrande, where preserving contextual connection between questions and answers together with removing the leakage of the correct answer through grammatical gender markers plays a crucial role in evaluation consistency.

Language	AVG Δ	Language	AVG Δ
Greek	+3.89%	Romanian	+2.09%
Ukrainian	+2.7%	Slovak	+1.66%
Turkish	+2.65%	Estonian	+1.63%
Lithuanian	+2.6%	Bulgarian	+1.37%

Table 4: Average improvement per language across all benchmarks

As shown, our proposed framework effectively enhances translation quality for multilingual benchmarks, leading to more reliable and accurate evaluations of language models across diverse languages.

5 Conclusion

We present a novel automated translation framework designed to enable rapid, high-quality translation of datasets and benchmarks while minimizing human intervention. Our integrated methods demonstrate substantial improvements on WMT24++ and FLORES benchmarks as measured by COMET scores, with T-RANK and USI achieving the strongest performance through test-time compute strategies. The quality of our translated benchmarks is further validated through COMET Quality Estimation scores and LLM-as-a-judge evaluations, which confirm meaningful improvements over existing translations. Notably, evaluation results of mid-sized models such as Gemma 3, Qwen 3 and Llama 3.1 on our improved translations yield higher scores compared to existing benchmark translations, providing additional evidence of enhanced translation quality. These findings demonstrate that our framework effectively balances translation accuracy, computational efficiency, and scalability, offering a practical solution for creating high-quality multilingual evaluation resources. Future work should explore adaptive method selection based on translation difficulty, integration of dedicated quality models, and comprehensive evaluation across open-weight models to further enhance the framework's capabilities on languages beyond Europe.

Limitations

Our study has several limitations that warrant consideration. First, we employ LLM-based scoring rather than dedicated translation quality models like COMET for the Best-of-N selection process, which may affect the reliability of candidate ranking. Second, our approach applies uniform methods across all entries without automatically estimating translation difficulty per input, though adaptive method selection based on text complexity could potentially improve efficiency and quality. Machine translation benchmarks demonstrate that the most advanced and computationally expensive methods do not always yield superior results for shorter text sequences, particularly those not used in question-answering contexts. We defer to users of our framework to select methods based on their specific objectives and resource constraints. Third, we rely primarily on closed-source models for translation, with limited testing of open-weight alternatives. We hypothesize that our proposed methods would yield greater benefits for open-weight models, which typically demonstrate weaker performance in zero-shot translation settings, though this requires validation through comprehensive evaluation. Finally, while we focus on Eastern and Southern European languages, further research is needed to assess the generalizability of our framework across a broader range of low-resource languages with diverse linguistic characteristics.

Ethics Statement

This work aims to improve multilingual benchmark quality to enable more equitable evaluation of language models across diverse languages. We acknowledge several ethical considerations. First, our reliance on closed-source models raises reproducibility concerns; we plan to extend evaluation to open-weight alternatives to promote accessibility. Second, while focusing on Eastern European languages addresses an important gap, many lower-resource languages remain underrepresented, and we encourage extending these methods to broader linguistic contexts. Third, we recognize that improved benchmarks may still be imperfect or subject to translation model bias; however, accurate multilingual evaluation is crucial for developing culturally aware and safer AI systems. Our automated translation approach avoids potential labor exploitation concerns associated with human translation.

The code and translated benchmarks produced in this work will be released under permissible open licenses to enable community validation and reproducibility. Generative AI systems were employed exclusively for language assistance, including paraphrasing, spell-checking, and stylistic refinement of the authors' original content, as well as for generating boilerplate code. All core research contributions, experimental design, and findings represent the original work of the authors. Used benchmarks and datasets are publicly available under permissive licensing; any ethical considerations related to their use are discussed in the original publications.

References

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,

Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Preprint*, arXiv:2310.10482.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111. Association for Computational Linguistics.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *Preprint*, arXiv:2506.19468.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Ammar Khairi, Daniel D'souza, Marzieh Fadaee, and Julia Kreutzer. 2025a. Making, not taking, the best of n. *Preprint*, arXiv:2510.00931.

Ammar Khairi, Daniel D'souza, Ye Shen, Julia Kreutzer, and Sara Hooker. 2025b. When life gives you samples: The benefits of scaling up inference compute for multilingual llms. *Preprint*, arXiv:2506.20544.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *Preprint*, arXiv:2309.11925.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Lei Tang, Jinghui Qin, Wenxuan Ye, Hao Tan, and Zhijing Yang. 2025. Adaptive few-shot prompting for machine translation with pre-trained language models.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Translation Examples

In this subsection we will provide some translation examples showcasing common errors and how our

proposed methods address them. Figure 3 demonstrates how T-RANK identifies and corrects translation errors through competitive ranking. Unlike all other methods, we observe improved ability of the judge model to notice translation flaws when evaluating the translation in competitive manner. Table 6 illustrates a common issue where translating questions and answers separately can leak the correct answer through grammatical gender markers, and shows how the correct approach preserves context during translation. We also identify systematic translation errors present in existing multilingual benchmark translations for MMLU and Hellaswag, complementing the Winogrande example shown in Table 6. These errors arise when questions and answer options are translated independently, without shared grammatical context.

Table 7 documents four categories of errors found in MMLU translations produced by Global-MMLU and Okapi. Table 8 shows a concrete Hellaswag example where translating the question and answer options without the shared narrative context produces unnatural phrasing, grammatical errors, and broken cohesion between the question stem and the answer options.

Table 5 reports the positional bias observed during the T-RANK ranking step on MMLU (EN→UK). Even though the multi-round strategy rotates candidates across positions to mitigate bias, a residual preference for the candidate presented at input position 2 remains visible: it receives the lowest (best) average rank and holds rank 1 in all five most frequent rank combinations.

A.2 Benchmark Translation Statistics

In this subsection we provide statistics of selected benchmark translation, used splits and total number of samples used. We also provide information on models and methods used for final translation.

A.3 Detailed Evaluation Results

In this section we provide additional evaluation results on our own and existing translated benchmarks for analyzed languages. Tables 11 through 19 present performance comparisons across Ukrainian, Romanian, Slovak, Lithuanian, Bulgarian, Turkish, Greek and Estonian languages respectively, showing consistent improvements of our translations over existing benchmarks. Where available, we compare against Okapi/MuBench/Global-MMLU translations as baselines. For Bulgarian language we also compare against our older translation done with the help of professional translators. With these results we confirm, that translation quality influences the evaluation results across various models with some exceptions to Winogrande. We note that professionally translated Winogrande to Bulgarian shows better results than our automated translation, which indicates that for some languages and benchmarks human intervention is still required to achieve the best quality, though we believe that with our methods the need for such intervention is significantly reduced.

A.4 Machine Translation Benchmark Evaluation Results

In this section, we test GPT-4o-mini and Gemini-2.0-Flash model on WMT and FLORES translation benchmarks, using both reference-based and reference-free COMET evaluation setups. As mentioned before, for reference-based evaluation we use Unbabel/XCOMET-XL model and for reference-free QE evaluation we use Unbabel/wmt23-cometkiwi-da-xl model. Tables 20 through 23 present GPT-4o-mini results, while Tables 24 and 25 show Gemini-2.0-Flash performance.

We see, that on both datasets and evaluation types, USI method gives the best results. We also see, that using sampling single candidate from various prompts in both English and target language (2 English and 2 target languages prompts) gives additional boost in quality over single-prompt generation.

In this case, for Gemini-2.0-Flash we observe that T-RANK multi-prompt method can outperform USI on most language pairs, although the difference is quite small. We believe that both methods have their strong and weak points – while for datasets with shorter texts we recommends using USI, for benchmark translation we observed improved quality from using T-RANK method specifically, as shown in first subsection of appendix. Moreover, for Gemini-2.0-Flash model the ranking capabilities seem to be stronger, which may also contribute to better performance of T-RANK method. This supports our idea, that different methods may be more suitable for different models and use-cases.

A.5 Translation Prompts

Table 26 provides the complete prompt templates used for all described translation methods. Fig-

Translation Ranking (T-RANK) Correction Example

Candidate 1

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий життєвий цикл для свого розміру тіла. Що з наступного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в новозаселеному середовищі.
2. Це водний організм.
3. Він потребує відносно високої батьківської турботи про потомство.
4. Вік, в якому потомство саме може народжувати, є відносно молодим.

Candidate 2

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в новозаселеному середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської турботи про потомство.
4. Вік, в якому потомство саме може народжувати, є відносно молодим.

Candidate 3

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки за потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Selected Translation

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки за потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Reasoning

The final selected translation is generally accurate and conveys the original meaning well. However, there are minor improvements needed in terms of word choice and grammatical structure. The phrase 'відносно тривалий термін життя' could be simplified to 'відносно довгий термін життя' for better clarity. Additionally, the phrase 'батьківської опіки за потомством' can be improved to 'батьківської опіки над потомством' for grammatical correctness. Overall, the translation is good but requires these minor adjustments for optimal clarity and correctness.

Final Corrected Translation

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно довгий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки над потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Figure 3: Translation Ranking (T-RANK) Correction Example

ure 5 provides the prompt for LLM-as-a-judge to compare our translations versus Global-MMLU as described in Table 2.

Universal Self-Improvement (USI) Correction Example

Candidate 1

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий життєвий цикл для свого розміру тіла. Що з наступного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в новозаселеному середовищі.
2. Це водний організм.
3. Він потребує відносно високої батьківської турботи про потомство.
4. Вік, в якому потомство саме може народжувати, є відносно молодим.

Candidate 2

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в новозаселеному середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської турботи про потомство.
4. Вік, в якому потомство саме може народжувати, є відносно молодим.

Candidate 3

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки за потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Selected Translation

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий термін життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки за потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Reasoning

<no reasoning used for USI method>

Final Corrected Translation

Question: Ви спостерігаєте вид, який народжує лише одне потомство за раз і має відносно тривалий цикл життя для свого розміру. Що з наведеного, ймовірно, також вірно для цього організму?

Answers:
1. Він живе в нещодавно колонізованому середовищі.
2. Це водний організм.
3. Він потребує відносно великої батьківської опіки за потомством.
4. Вік, в якому потомство може самостійно народжувати, відносно молодий.

Figure 4: Universal Self-Improvement (USI) Correction Example

Input position	Avg. rank	Top-5 rank combination (positions 1–5)	Occurrences
1	3.01	(4, 1, 3, 2, 5)	4731
2	**2.06**	(3, 1, 2, 4, 5)	3141
3	2.93	(4, 1, 2, 3, 5)	2997
4	3.07	(3, 1, 4, 2, 5)	2600
5	3.93	(2, 1, 3, 4, 5)	2428
Equal ranks			6324

Table 5: Positional bias in T-RANK ranking for MMLU (EN→UK, n=5 candidates for p=1, translation model *gpt-4o-mini*). Average rank per input position (lower = ranked better) and the five most frequent rank combinations. Position 2 receives the best average rank (2.06) and holds rank 1 in every top combination, confirming a residual bias towards the second candidate despite the multi-round rotation strategy.

✗**Bad Example (MuBench, Answer Leakage)**	✓**Good Example (Ours, Correct Translation)**
Вони хвилювалися, що вино зіпсує ліжко та ковдру, але _ не було зіпсовано.	Вони хвилювалися, що вино зіпсує ліжко та ковдру, але _ не бу(-в/-ла/-ло/-ли) зіпсовано.
What does the blank _ refer to?	What does the blank _ refer to?
Option A: ковдра Option B: ліжко	Option A: ковдра Option B: ліжко
Answer with A or B.	Answer with A or B.
Answer: ліжко	Answer: ліжко ?

Table 6: Winogrande translation examples. Left (MuBench UK Winogrande): answer leakage where the correct answer "ліжко" is implied grammatically in the question itself. Right (our translation): correct translation with verb morphology masking "бу(-в/-ла/-ло/-ли)" to prevent answer leakage.

Issue Type	Example	Impact
Semantic Drift	**Original:** "relatively long lifespan" **Translated:** "життєвий цикл" (life cycle) **Should be:** "тривалість життя" (lifespan)	Changes the question meaning; "life cycle" refers to developmental stages, not longevity
Wrong Terminology	**Original:** "aquatic organism" **Translated:** "водяний організм" **Should be:** "водний організм"	"Водяний" means "watery" (like soup), while "водний" is the correct scientific term for aquatic
Grammar Errors	**Original:** "parental care for offspring" **Translated:** "турботи за потомством" **Should be:** "турботи про потомство"	Incorrect preposition usage creates unnatural phrasing that may confuse native speakers
Literal Translation	**Original:** "Only I" **Translated:** "Тільки я" **Should be:** "Тільки I"	The Latin numeral "I" is mis-read as the letter and rendered as the Ukrainian pronoun "я" (I/me); the Roman numeral must be preserved in Latin script

Table 7: MMLU translation errors (Ukrainian) found in existing benchmarks (Global-MMLU, Okapi). Four recurring error categories with concrete examples and their impact on evaluation reliability.

✗ Bad Example (Okapi, Literal Translation)	✓ Good Example (Ours, Natural Phrasing)
Ми бачимо початковий екран заголовка. Ми бачимо дівчинку біжать та робить високий стрибок, пройшовши через перекладину. Ми Яке закінчення має найбільший сенс? A) бачимо близько 30 дівчат стрибають. B) бачимо, як вона закінчує перекладину та виконує рутину. C) потім ми бачимо повтор і повільний повтор. D) бачимо, як дівчина кидає білий м'яч, і він летить прямо по полю. Відповідь: C	Ми бачимо титульний екран. Ми бачимо, як дівчина біжить, виконує стрибок у висоту і перелазить через планку. Ми Яке закінчення має найбільший сенс? A) бачимо, як близько 30 дівчат стрибають. B) бачимо, як вона закінчує стрибок над планкою і забирає свою рутину. C) потім бачимо повтор і уповільнений повтор. D) бачимо, як дівчина кидає білий м'яч, і він летить прямо через поле. Відповідь: C
Issue: *"дівчинку біжать" mixes singular accusative with a plural verb; "пройшовши через перекладину" is an awkward circumlocution for "making it over the bar"; option C inserts a disconnected subject "ми" that breaks narrative flow.*	***Improvement:*** *"дівчина біжить" restores correct agreement; "перелазить через планку" is the natural sport phrase; "потім бачимо" preserves the* "Ми. . . потім бачимо" *narrative flow.*

Table 8: Hellaswag translation example (Ukrainian). Original English: "We see an opening title screen. We see a girl run and perform a high jump and make it over the bar. We. . . — which ending makes the most sense?"

Benchmark	Train	Val	Dev	Test	All
MMLU	–	1531	285	14042	15858
Winogrande	–	1267	–	–	–
ARC-Challenge	1119	299	–	1172	2291
Hellaswag	–	–	–	10042	10042
Total	1119	3097	285	25556	**29757**

Table 9: Number of examples in each benchmark split (– indicates split not available or not used).

Benchmark	UK	SK	RO	LT	ET	BG	EL	TR
MMLU	4o/TR	4o/TR	4o/TR	4o/USI	4o/USI	4o/TR	4o/USI	4o/USI
Hellaswag	4o	4o	4o	G2F	G2F	G2F	G2F	G2F
ARC	4o	G2F	G2F	G2F	G2F	G2F	G2F	G2F
Winogrande	4o	G2F	G2F	G2F	G2F	G2F	G2F	G2F

Table 10: Model and method for benchmark translation. 4o = GPT-4o-mini, G2F = Gemini-2.0-Flash, TR = T-RANK. If method not specified, USI was used.

	Hellaswag			Winogrande			ARC-Challenge			MMLU		
Model	O	Ot	Δ%	O	Ot	Δ%	O	Ot	Δ%	O	Ot	Δ%
Gemma-3-12B-IT	0.687	0.625	+6.2%	0.580	0.619	-3.9%	0.517	0.470	+4.8%	0.614	0.603	+1.2%
Llama-3.1-8B	0.570	0.540	+3%	0.549	0.495	+5.5%	0.431	0.416	+1.5%	0.497	0.489	+0.9%
Gemma-3-4B-IT	0.578	0.528	+5%	0.540	0.501	+3.9%	0.453	0.417	+3.7%	0.453	0.444	+1%
Qwen3-8B-IT	0.540	0.520	+2%	0.569	0.546	+2.3%	0.448	0.404	+4.4%	0.619	0.599	+2%

Table 11: Ukrainian. O=Ours, Ot=Other (Okapi/MuBench/Global-MMLU)

Model	Hellaswag			Winogrande			ARC-Challenge			MMLU		
	O	Ot	Δ%	O	Ot	Δ%	O	Ot	Δ%	O	Ot	Δ%
Gemma-3-12B-IT	0.678	0.681	-0.2%	0.661	0.589	+7.2%	0.486	0.454	+3.2%	0.628	0.614	+1.4%
Gemma-3-4B-IT	0.567	0.567	0%	0.575	0.531	+4.5%	0.412	0.394	+1.8%	0.479	0.473	+0.6%
Llama-3.1-8B-IT	0.576	0.581	-0.5%	0.616	0.521	+9.5%	0.358	0.357	+0.2%	0.529	0.519	+0.9%
Qwen3-8B-IT	0.538	0.537	+0.1%	0.588	0.566	+2.3%	0.368	0.359	+0.8%	0.650	0.632	+1.8%

Table 12: Romanian. O=Ours, Ot=Other (Okapi/MuBench/Global-MMLU)

Model	Winogrande			ARC-Challenge		
	Ours	Other	Δ%	Ours	Other	Δ%
Gemma-3-12B-IT	0.610	0.589	+2.1%	0.539	0.509	+3%
Gemma-3-4B-IT	0.554	0.517	+3.7%	0.448	0.444	+0.4%
Llama-3.1-8B-IT	0.547	0.515	+3.2%	0.443	0.403	+4%
Qwen3-8B-IT	0.555	0.559	-0.4%	0.428	0.427	+0.1%

Table 13: Slovak. Other=MuBench/Okapi

Model	Winogrande			MMLU		
	Ours	Other	Δ%	Ours	Other	Δ%
Gemma-3-12B-IT	0.631	0.557	+7.3%	0.585	0.573	+1.3%
Gemma-3-4B-IT	0.530	0.506	+2.3%	0.433	0.410	+2.3%
Llama-3.1-8B-IT	0.527	0.498	+2.9%	0.431	0.417	+1.4%
Qwen3-8B-IT	0.551	0.530	+2.1%	0.553	0.541	+1.2%

Table 14: Lithuanian. Other=MuBench/Global-MMLU

Model	Winogrande		
	Ours	Other	Δ%
Gemma-3-12B-IT	0.615	0.566	+4.9%
Gemma-3-4B-IT	0.541	0.516	+2.5%
Llama-3.1-8B-IT	0.543	0.509	+3.4%
Qwen3-8B-IT	0.517	0.503	+1.4%

Table 15: Estonian. Other=MuBench

Model	Winogrande			MMLU		
	Ours	Other	Δ%	Ours	Other	Δ%
Gemma-3-12B-IT	0.635	0.571	+6.5%	0.587	0.581	+0.7%
Gemma-3-4B-IT	0.583	0.514	+7%	0.447	0.437	+1%
Llama-3.1-8B-IT	0.573	0.518	+5.5%	0.494	0.480	+1.4%
Qwen3-8B-IT	0.545	0.571	-2.7%	0.588	0.569	+1.8%

Table 16: Turkish. Other=MuBench/Global-MMLU

Model	Winogrande			MMLU		
	Ours	Other	Δ%	Ours	Other	Δ%
Gemma-3-12B-IT	0.657	0.601	+5.6%	0.593	0.580	+1.3%
Gemma-3-4B-IT	0.598	0.518	+8.0%	0.428	0.421	+0.7%
Llama-3.1-8B-IT	0.593	0.520	+7.3%	0.465	0.446	+2%
Qwen3-8B-IT	0.581	0.528	+5.3%	0.563	0.553	+1.1%

Table 17: Greek. Other=MuBench/Global-MMLU

Model	Hellaswag			Winogrande			ARC-Challenge			MMLU		
	Ours	Other	Δ%	Ours	Other	Δ%	Ours	Other	Δ%	Ours	Other	Δ%
Gemma-3-12B-IT	0.708	0.690	+1.8%	0.643	0.677	-3.5%	0.526	0.495	+3.1%	0.605	0.619	-1.3%
Gemma-3-4B-IT	0.590	0.569	+2.2%	0.588	0.595	-0.7%	0.430	0.410	+2%	0.454	0.469	-1.5%
Llama-3.1-8B-IT	0.539	0.528	+1.1%	0.575	0.605	-3.1%	0.394	0.367	+2.7%	0.481	0.487	-0.6%
Qwen3-8B-IT	0.536	0.518	+1.8%	0.602	0.629	-2.7%	0.414	0.394	+2%	0.619	0.617	+0.2%

Table 18: Bulgarian. Other=INSAIT (professional translators)

Model	Winogrande		
	Ours	Other	Δ%
Gemma-3-12B-IT	0.643	0.597	+4.6%
Gemma-3-4B-IT	0.588	0.506	+8.2%
Llama-3.1-8B-IT	0.575	0.505	+6.9%
Qwen3-8B-IT	0.602	0.559	+4.3%

Table 19: Bulgarian. Other=MuBench

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.726	0.741	0.882	0.741	0.788	0.751	0.718	0.802
SC	0.708	0.725	0.869	0.735	0.788	0.749	0.715	0.793
BoN n=5	0.743	0.756	0.895	0.767	0.809	0.788	0.736	0.819
USI n=5	0.750	0.759	0.899	0.766	0.806	0.791	0.733	0.817
T-RANK n=5	0.739	0.745	0.885	0.749	0.799	0.776	0.729	0.811
USI multi-prompt p=5	**0.755**	**0.764**	**0.898**	**0.771**	**0.809**	**0.793**	**0.736**	**0.825**
T-RANK multi-prompt p=5	0.742	0.756	0.883	0.753	0.797	0.778	0.726	0.811

Table 20: WMT QE (Quality Estimation) reference-free COMET scores for GPT-4o-mini

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.827	0.822	0.873	0.788	0.821	0.834	0.776	0.820
SC	0.821	0.817	0.869	0.790	0.822	0.834	0.834	0.821
BoN n=5	0.844	0.837	0.884	0.805	0.838	0.852	0.791	0.836
USI n=5	0.843	0.839	0.888	0.807	0.842	0.856	0.791	0.838
T-RANK n=5	0.840	0.834	0.885	0.803	0.836	0.855	0.793	0.837
USI multi-prompt p=5	**0.849**	**0.847**	**0.891**	**0.817**	**0.848**	0.862	**0.803**	**0.844**
T-RANK multi-prompt p=5	0.845	0.841	0.887	0.812	0.843	**0.862**	0.798	0.841

Table 21: WMT COMET reference-based scores for GPT-4o-mini

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.904	0.906	0.947	0.906	0.927	0.903	0.904	0.896
SC	0.902	0.908	0.950	0.911	0.935	0.902	0.903	0.896
BoN n=5	0.910	0.917	0.956	0.921	0.947	0.916	0.912	0.911
USI n=5	0.913	0.915	0.956	0.924	0.943	0.917	0.910	0.909
T-RANK n=5	0.899	0.906	0.952	0.909	0.932	0.911	0.903	0.902
USI multi-prompt p=5	**0.919**	**0.921**	**0.961**	**0.932**	**0.948**	**0.922**	**0.918**	**0.912**
T-RANK multi-prompt p=5	0.905	0.912	0.954	0.912	0.935	0.912	0.908	0.904

Table 22: FLORES QE (Quality Estimation) reference-free COMET scores for GPT-4o-mini

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.937	0.938	0.939	0.906	0.894	0.943	0.920	0.920
SC	0.937	0.937	0.936	0.911	0.902	0.940	0.916	0.923
BoN n=5	0.943	0.945	0.945	0.917	0.918	0.951	0.927	0.931
USI n=5	0.945	0.942	0.947	0.922	0.919	0.951	0.924	0.932
T-RANK n=5	0.938	0.937	0.943	0.910	0.903	0.947	0.920	0.929
USI multi-prompt p=5	**0.947**	**0.946**	**0.949**	**0.928**	**0.918**	**0.952**	**0.931**	**0.934**
T-RANK multi-prompt p=5	0.940	0.943	0.944	0.915	0.907	0.950	0.925	0.932

Table 23: FLORES COMET reference-based scores for GPT-4o-mini

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.688	0.703	0.777	0.704	0.737	0.708	0.683	0.719
SC	0.687	0.706	0.777	0.706	0.736	0.709	0.685	0.722
BoN n=5	0.687	0.707	0.780	0.710	0.743	0.713	0.689	0.723
USI multi-prompt p=5	**0.698**	**0.716**	**0.785**	**0.718**	**0.753**	**0.721**	**0.697**	**0.731**
T-RANK multi-prompt p=5	0.697	0.713	0.781	0.716	0.749	0.718	0.695	0.726

Table 24: WMT QE (Quality Estimation) reference-free COMET scores for Gemini-2.0-flash

Method	EN→UK	EN→SK	EN→RO	EN→LT	EN→ET	EN→BG	EN→TR	EN→EL
Baseline	0.828	0.829	0.867	0.805	0.847	0.840	0.778	0.821
SC	0.826	0.834	0.868	0.808	0.845	0.841	0.779	0.824
BoN n=5	0.828	0.836	0.870	0.814	0.852	0.843	0.785	0.823
USI multi-prompt p=5	**0.841**	0.843	0.881	0.826	**0.862**	0.856	0.797	0.836
T-RANK multi-prompt p=5	0.841	**0.849**	**0.882**	**0.827**	0.861	**0.859**	**0.800**	**0.838**

Table 25: WMT COMET reference-based scores for Gemini-2.0-flash

Method	Prompt Template
Base Translation	**Instructions:** Imagine you're part of a team at an international education center that's revamping its exams for a global audience. Your job is to translate an English question and its answer options into <target_language> so that students from <target_language> schools can be evaluated too. Just provide the final translation—leave out any extra comments or explanations. Use language which is authentic for <target_language> natives. Remember to keep the answer options connected to the question, using the same format as the original (a list for multiple choices or plain text for a single answer). Please do not translate valid code in any of the programming languages. **Original text:** {"Original_question": "<original_question>", "Original_answers": "<original_answers>"} **Output instructions:** Now, please give your final translation in <target_language> exactly in this format, with only the translated content: {"Question": "your_translated_question", "Answers": "translated_answers"}
USI Judge	My task is to translate BENCHMARK questions with answers from English to <target_language>. Your task is to evaluate if the response preserves the original question idea and to verify the correctness of declension and conjunction of words in the target language. The original text in English is: Question: <original_question>, Answers: <original_answers> I have generated the following responses: <responses> Combine the best features from responses to form the best response from grammatical and coherent points of view. Look for the next metrics: * Quality of translation, including grammatical correctness * Domain knowledge - were the terms correctly translated w.r.t to domain? Were coding terms or function names preserved (not translated)? * Is the question text fully translated? Is the question idea preserved? * Are all the answer options/texts fully translated? * Are the words written correctly? Are there any typos? Output only the selected response: Question: selected question, Answer: selected answers
T-RANK Ranking	My task is to translate BENCHMARK questions from English to <target_language>. Your task is to rank my translations and select the best one. **Ranking criteria:** * Quality of translation * Domain knowledge - were the terms correctly translated w.r.t to domain? Were coding terms or function names preserved (one would usually not translate coding operators or function names)? * Is the question text fully translated? Is the question idea preserved? * Are all the answer options/texts fully translated? * Are the words written correctly? Are there any typos? * Are the declension and the conjunction of words written correctly in the target language if you connect answer options with the question? **Original:** {"original_question": "<original_question>", "original_answers": <original_answers>} **Candidates:** <responses> **Instruction:** Select the best response (1st place = best). Correct if needed before output. **Output:** Reasoning, then: {"summary": "...", "final_ranks": {...}, "rankings_list": [...], "best_translation": {...}}
Best-of-N Scoring	My task is to translate questions from English to <target_language>. Score my translations 1-10 (10 = best). **Original**: Question: <original_question>, Answers: <original_answers> **Scoring metrics**: (1) Translation quality; (2) Question fully translated?; (3) All answers translated?; (4) Question idea preserved?; (5) Correct grammar? **Responses**: <responses> Output scores only: Response 1: score, Response 2: score, ..., Answers: [list of scores]

Table 26: Translation prompt templates for different methods

LLM-as-a-judge Quality Evaluation Prompt

Current User Query
"instruction": Your task is compare two translation from from English to Ukrainian and compare their advantages and disadvantages. Lastly, you have to pick one of three choices: A is better, B is better or they are equal.
Original text (English)
<original_text>
Translation A (Ukrainian)
<begin_of_translation_A> <output_1> <end_of_translation_A>
Translation B (Ukrainian)
<begin_of_translation_B> <output_2> <end_of_translation_B>
Evaluation
Rules
You should compare the above two translations from English to Ukrainian based on your analysis of the user queries. You should first write down your analysis and the checklist that you used for the evaluation, and then provide your assessment according to the checklist. There are two choices to give your final assessment: ["A+", "B+",], which correspond to the following meanings:
- 'A+': Translation A is better than Translation B.
- 'B+': Translation B is better than Translation A.
- 'T=': Translation A and Translation B are equally good. Tie.
Output Format
First, please output your analysis for each model translation from English to Ukrainian, and then summarize your assessment to three aspects: "reason A>B", and "reason B>A", and finally make your choice for the final assessment. Please provide your evaluation results in the following json format (starting with and ending with), with no quotes around the curly brackets by filling in the placeholders in []:
analysis_of_A: <analysis of Translation A>,
analysis_of_B: <analysis of Translation B>,
reason_of_A_equals_B: <where Translation A and B perform equally well>,
reason_of_A_better_than_B: <where Translation A is better than Translation B>,
reason_of_B_better_than_A: <where Translation B is better than Translation A>,
choice: <A+, B+ or T=>
Your translation should be a valid json dictionary following exactly this format.
Your answer:

Figure 5: LLM-as-a-judge Quality Evaluation Prompt