Towards Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset (GIST)

Jiarui Liu,[1*] Iman Ouzzani,[1*] Wenkai Li,[1*] Lechen Zhang,[2] Tianyue Ou,[1] Houda Bouamor,[1] Zhijing Jin,[3] Mona Diab[1]
[1]Carnegie Mellon University,
[2]University of Michigan, [3]University of Toronto
{jiarui15, iouzzani, wenkail}@andrew.cmu.edu,
zjin@cs.toronto.edu, mdiab@andrew.cmu.edu

Abstract

The field of machine translation has achieved significant advancements, yet domain-specific terminology translation, particularly in AI, remains challenging. We introduce GIST, a large-scale multilingual AI terminology dataset containing 5K terms extracted from top AI conference papers spanning 2000 to 2023. The terms are translated into Arabic, Chinese, French, Japanese, and Russian using a hybrid framework that combines LLMs for extraction with human expertise for translation. The dataset's quality is benchmarked against existing resources, demonstrating superior translation accuracy through crowdsourced evaluation. GIST is integrated into translation workflows using post-translation refinement methods that require no retraining, where LLM prompting consistently improves BLEU, COMET, and other scores. A web demonstration on the ACL Anthology platform highlights its practical application, showcasing improved accessibility for non-English speakers. This work aims to address critical gaps in AI terminology resources and fosters global inclusivity and collaboration in AI research.

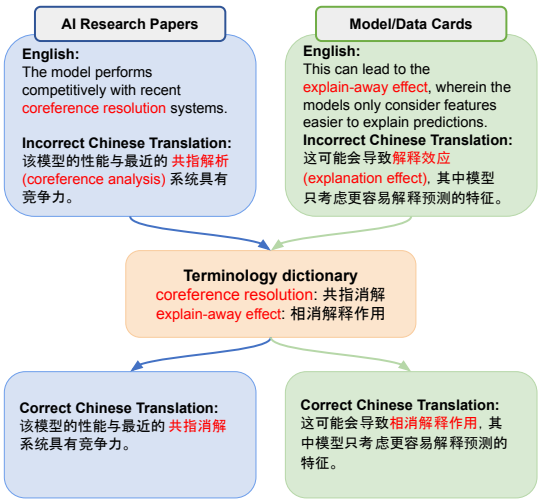

Figure 1: Direct translations of AI research papers and model cards using Google Translate generally offer fair quality but often fail to accurately translate AI-specific terminologies, potentially causing confusion or misunderstanding for readers. Our work addresses this issue by providing high-quality translations for a wide range of such terms, which can be efficiently integrated post-hoc to enhance the initial translations.

1 Introduction

The field of machine translation has made significant progress, with state-of-the-art models excelling across diverse tasks (Brown et al., 1990; Wu et al., 2016; Goyal et al., 2022; Haddow et al., 2022) and demonstrating effectiveness in translating between high-resource and low-resource languages (Yao and Wan, 2020; Costa-jussà et al., 2022; Ranathunga et al., 2023). Despite these successes, translating domain-specific scientific texts remains a persistent challenge, particularly for terminology translation (Cabré, 2010; Shuttleworth, 2014; Naveen and Trojovskỳ, 2024). General-purpose translation systems often falter in accurately translating specialized terminology, leading to loss of critical details (Dagan and Church, 1994; Haque et al., 2020a), or worse yet possibly leading to misinterpretations (Chmutina et al., 2021; Yue et al., 2024).

In the field of AI, terms such as "Coreference Resolution" or "Explain-Away Effect," are frequently translated incorrectly or inconsistently, undermining comprehension for global researchers and practitioners (Khuwaileh and Khwaileh, 2011; Tehseen et al., 2018), as illustrated in Figure 1. For non-English readers, who comprise a substantial portion of the global population both within and beyond the AI community (Ammon, 2003; Ding et al., 2023), such inaccuracies in AI research papers and blog translation hinder access to essential knowledge, stifling research innovation and collaboration across linguistic boundaries (Amano et al.,

[*]Equal contribution.

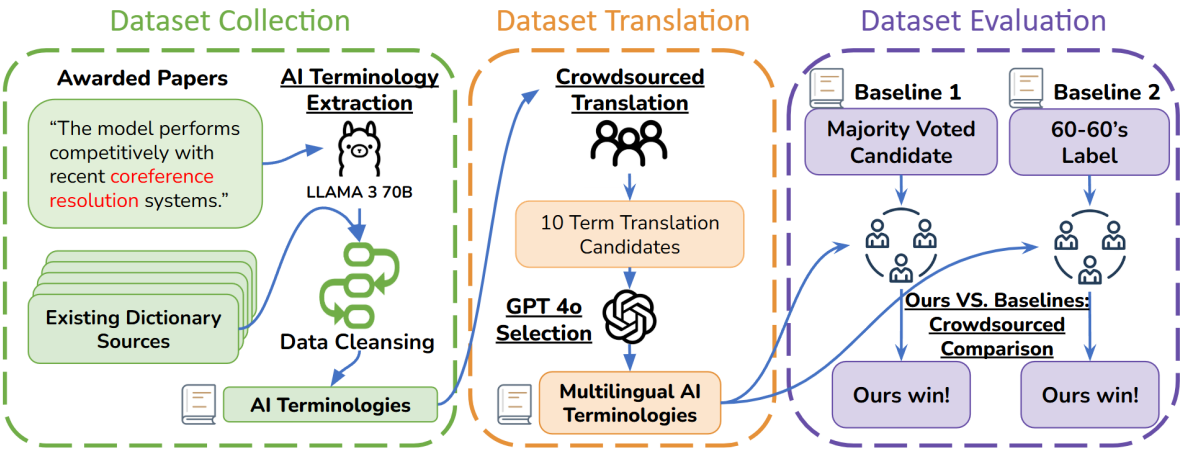

Figure 2: Overview of GIST creation. AI terminology are extracted from awarded papers using an LLM and then combined with existing terminology dictionaries. After data cleansing, translations into five languages are generated via crowdsourcing, and an LLM is used to select the best candidate translation. Dataset quality is evaluated against two baselines: majority-vote-selected translations and translations from the 60-60 evaluation set, demonstrating the superior quality of GIST.

2023; Bahji et al., 2023). The implications extend to AI resource documentation, such as model and data cards hosted on platforms such as Hugging Face,[1] where errors in translated terminology can lead to misunderstandings or incorrect usage of models and datasets. This highlights the urgent need for precise and standardized multilingual AI terminology resources to support equitable access to AI knowledge (Ahuja et al., 2023; Liu et al., 2024).[2]

Existing efforts to address this gap have relied heavily on manual curation of specialized terminology by domain experts, a process that is time-consuming, resource-intensive, and difficult to scale (Wang et al., 2013; Freitag et al., 2021). Meanwhile, large language models (LLMs) have shown promise in automated terminology translation (Feng et al., 2024), while their outputs often misalign with human expert standards (Zhang et al., 2024), and different models often yield inconsistent translations (Banik et al., 2019; Prieto Ramos, 2021). The ACL 60-60 initiative in 2022 curated and translated AI-specific terms, and showcased potential for multilingual scientific communication and AI inclusivity (Salesky et al., 2023).[3] However, its small scale and narrow scope underscore the need for a more comprehensive multilingual AI terminology resource at scale.

[1] huggingface.co/docs/hub/en/model-cards, huggingface.co/docs/hub/en/datasets-cards

[2] See Appendix A for a more in-depth discussion of our motivation.

[3] 2022.aclweb.org/dispecialinitiative.html

To address this, we introduce **G**lossary of Multilingual A**I** **S**cientific **T**erminology (GIST), the first large-scale multilingual AI terminology dataset, compiling 5K AI-specific terms from award-winning papers at 18 top-tier AI conferences (2000–2023). Using a hybrid approach combining LLMs for extraction and verification with human expertise for translation, we provide high-quality translations into Arabic, Chinese, French, Japanese, and Russian (Abel and Meyer, 2013).

To enhance accessibility of our work, we integrate the curated terminology into machine translation pipelines without requiring model retraining. We evaluate three post-translation refinement methods: prompting, word alignment and replacement, and constrained decoding. Our experiments reveal that the prompting method effectively integrates the curated dictionary, consistently improving translation quality as measured by BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), ChrF (Popović, 2015), ChrF++ (Popović, 2017), and TER (Snover et al., 2006). Additionally, we develop a website demonstration that incorporates our dataset into translations of ACL Anthology research papers, showcasing its practical application for non-English speakers in the AI field.

The contributions of this work are summarized as follows:

- We create GIST, the first large-scale multilingual AI terminology dataset, with 5K terms from 879 awarded top AI conference papers

(2000–2023) and broad domain coverage;[4]

- We develop an effective and efficient translation framework combining LLMs and human expertise to translate English terms into five languages, achieving high-quality results validated by automatic and human evaluations.
- To enhance accessibility, we explore three approaches to integrate the curated terminology into machine translation pipelines without retraining, and develop a website to showcase its practical use for AI paper translation.

By addressing the critical gap in multilingual AI terminology resources, this work contributes a robust solution to support equitable access to AI knowledge, fostering inclusivity and collaboration in the global research community.

2 Related Work

Multilingual Terminology Datasets The ACL 60-60 evaluation sets represent an important effort in translating multilingual terminology from AI presentations, but are limited to just over 250 terms (Salesky et al., 2023).

Scientific terminology datasets are typically built using two approaches. The first relies on human multilingual experts for extracting and translating terminology (Awadh, 2024; Kim et al., 2024). In cases of limited expert availability, human crowdsourcing with aggregation techniques has shown excellent results (Zaidan and Callison-Burch, 2011; Chan et al., 2023). The second approach uses ML and machine translation tools for automatic collection and translation, including log-likelihood comparisons (Haque et al., 2018), machine translation-based data synthesis (Haque et al., 2020b; Fernando et al., 2020; Manzini et al., 2022; Moslem et al., 2023), platform-based terminology linkers (Arcan et al., 2014), classifier training (Jin et al., 2013; Schumann and Martínez Alonso, 2018), and prompting LLMs (Nishio et al., 2024; Shamsabadi et al., 2024).

However, fully human-driven approaches can be costly for large-scale multilingual datasets, while fully automated ones often lack accuracy (Giguere, 2023). To address this, our framework integrates LLM-based extraction, human filtering, human translation, LLM validation, and merging with existing dictionaries, efficiently and effectively expanding existing terminology datasets.

[4]Our code and data is available at https://github.com/jiarui-liu/MultilingualAITerminology.

Integrating Domain Terminology into Machine Translation Integrating newly collected domain terminology into machine translation systems has led to a variety of research efforts. One common approach involves training methods, such as augmenting training data with input-output pairs that include terminology for supervised fine-tuning (Dinu et al., 2019; Niehues, 2021), or modifying model architectures to enhance terminology awareness (Dinu et al., 2019; Conia et al., 2024). However, these training-based approaches are inefficient for adapting to new terminology, limiting their real-world applicability.

To tackle this, terminology-aware decoding methods have emerged as a more flexible alternative. These methods, which include variants of constrained beam search (Anderson et al., 2016; Hokamp and Liu, 2017a; Chatterjee et al., 2017; Hasler et al., 2018; Post and Vilar, 2018), slightly compromise translation accuracy for greater adaptability. Another strategy is post-hoc editing of generated translations, which typically employs word alignment techniques to identify and replace term translations in the output (Zenkel et al., 2019; Chen et al., 2020; Ferrando et al., 2022). Recently, LLMs have been leveraged to integrate expected terminology directly into the translations (Bogoychev and Chen, 2023). Our work explores multiple terminology integration approaches, including constrained beam search, decoding logits adjustment, word alignment and replacement, and refinement through prompting.

3 GIST Dataset Construction

We construct GIST, a dataset comprising around 5K English AI terminology and their translations into Arabic, Chinese, French, Japanese, and Russian. The basic lexical statistics of the dataset is presented in Table 1. Additional dataset statistics are presented in Appendix B.1.

3.1 Terminology Curation

Our dataset follows the ACL 60-60 initiative that aims to collect scientific terminology in the AI field. We source AI terminology from two primary channels: AI research papers published online and existing AI terminology dictionaries. For the research papers, our objective is to compile a substantial number of terms from high-quality AI papers spanning a long time frame. To identify representative papers, a natural approach would be to crawl the

	Arabic	Chinese	French	Japanese	Russian
# Terms	4,844	6,426	6,527	4,770	5,167
Unique En Words	2,470	3,244	3,470	2,424	2,615
Unique Tgt Words	3,161	2,838	4,036	2,050	4,210
En Words/Term	2.02	2.05	2.07	2.02	2.01
Tgt Words/Term	2.36	2.26	2.68	2.53	2.16
En Chars/Term	16.99	17.26	17.44	16.96	16.94
Tgt Chars/Term	15.22	4.66	21.27	6.89	20.20

Table 1: Statistics of the dataset across languages. "En" denotes English, and "Tgt" denotes the target language. Statistics with standard deviations are presented in Table 4.

most cited papers using online search engines or platforms. However, no tools or APIs currently enable this. Instead, we focus on awarded papers announced on the websites of top AI conferences. Specifically, we collect all awarded papers, spanning awards such as Best Paper, Outstanding Paper, and other recognitions from the venues listed in Table 7, covering the years 2000 to 2023. This approach ensures comprehensive coverage of recognized research work. In total, we collect 879 paper PDFs from arXiv and other online repositories and process them into text files using SciPDF.[5] As we analyze later in Section 5, this collection strategy provides broad terminology and domain coverage of AI terminology.

As supported by previous research (Xu et al., 2024; Dagdelen et al., 2024; Shamsabadi et al., 2024), recent LLMs such as `LLaMA 3` (Dubey et al., 2024) have demonstrated strong capabilities in scientific terminology extraction tasks. While the definition of "AI" lacks a clear and universally agreed boundary, LLMs are trained on vast datasets that reflect human knowledge, enabling them to classify terms as AI-related based on their contextual relevance. Accordingly, we leverage `LLaMA-3-70B-Instruct` to extract AI terminology from award-winning papers, providing specific instructions to guide the extraction process.

To define AI-specific terminology, we impose the following criteria: (1) the term must be a noun or noun phrase, (2) it should be specialized to AI, encompassing core concepts, methods, models, algorithms, or systems, and (3) it should have either no meaning or a distinct meaning outside the AI domain. We process the text in sentence chunks of up to 64 words to stay within `LLaMA`'s optimal context-handling capabilities. Additionally, for each unique term, we record up to three different contexts during extraction to ensure sufficient contextual diversity.

After extraction, we perform multiple quality assurance steps. We remove terms that appear in only one paper to ensure representativeness. Moreover, we exclude abbreviations and terms starting with special characters, and filter out non-noun phrases. Duplicates are eliminated, and `GPT-4o` (Hurst et al., 2024) is employed to further refine the list by filtering out non-AI terms based on the same criteria. Finally, three AI domain experts review the terms to remove any remaining unqualified entries. To enhance the dataset, we also integrate terminology from external sources, including the 60-60 initiative dataset, government websites, Wikipedia, and other online resources. Consequently, the number of English terms varies across languages in GIST. Full details of the terminology collection process are provided in Appendix B.2.

3.2 Terminology Translation

In selecting target languages for translation, we aim to encompass a range of morphological complexities and varying levels of resource availability, particularly in AI-related publications. First, Chinese and Japanese exhibit minimal morphological variation in nouns and noun phrases, relying primarily on word order and context rather than inflection. In contrast, Arabic, French, and Russian are morphologically complex, characterized by extensive inflectional systems. Second, French, Chinese, Japanese, and Russian were selected due to the presence of large native-speaking scientific communities (Ammon, 2012; Chahal et al., 2022), which may have historically contributed to the development of well-established AI terminology. In contrast, Arabic, despite being widely spoken, may lack the same depth of scientific vocabulary, particularly in rapidly evolving fields such as AI. This disparity underscores the potential for our approach to contribute to the standardization and development of cohesive terminology across different languages in AI publications.

To evaluate whether state-of-the-art LLMs can effectively perform terminology translation as a generation task, we initially experiment with two advanced LLMs and one API: `Claude 3 Sonnet` (Anthropic, 2024), `GPT-3.5-Turbo` (OpenAI, 2023), and Google Translate API (Wu et al., 2016). We measure the agreement among the three methods using exact match. However, as shown in

[5] https://github.com/titipata/scipdf_parser

Table 8, the three-model agreement ratio was only around 15% for most languages except Chinese, and the two-model agreement ratio was only about 40%. These results reveal significant inconsistencies in the AI terminology translations produced by these systems, and highlight the need for human input to ensure reliable AI terminology translations. The detailed procedure and the prompt used for these translations is presented in Appendix B.3.

Given these findings, we opt for human annotation to ensure translation accuracy. To achieve this, we utilize Amazon Mechanical Turk (MTurk) for crowdsourced annotations. A demonstration of the MTurk task is shown in Figure 7. We instruct participants to take on the task only if they specialized in AI and are fluent in both English and one of the target languages. Annotators are tasked with generating accurate translations for each AI terminology, with relevant contexts provided and the terminology highlighted in yellow in context. To maintain quality, we implement a rigorous qualification process. Annotators are first tested on a toy set of 10 carefully selected AI terms, and only those who perform well are allowed to proceed with the full task. Additionally, we monitor submissions daily and filtered out participants who provide random or low-quality translations during the annotation process. All in all, for each term, we collect 10 annotations per target language.

Finally, we use `GPT-4o` to select the best translation from the annotators' submissions and Google API Translation for each term, ensuring high-quality results for our final dataset. As analyzed in Section 4.1, leveraging `GPT-4o` is crucial for maintaining the quality of the translations.

4 Dataset Quality Assessments

To thoroughly evaluate the translation quality in GIST, we conduct two additional crowdsourced rating tasks: (1) In Section 4.1, we investigate whether using `GPT-4o` is necessary to select the best translation candidate; (2) In Section 4.2, we compare the quality of our translation with the evaluation set from the 60-60 initiative.

4.1 Task 1: Is an LLM Necessary for Selecting the Best Translation Candidate?

To ensure the selection of the best candidate from the annotators' generations, we explore two methods for candidate selection. The first method only uses `GPT-4o` to select the best translation candidate among 10 annotations and one Google Translation for all terms. The second method relies on majority voting among the 11 translations: a translation candidate is selected if it appears in more than 5 out of 11 annotations. In cases where no majority is reached, `GPT-4o` is prompted to select the best translation candidate.

To evaluate these approaches, we conduct a separate MTurk task, involving a different group of participants, to compare the two methods on a randomly sampled subset of approximately 200 terms per language. Participants are asked to choose one of four options: (A) Both translations are good; (B) Method 1 translation is better; (C) Method 2 translation is better; (D) Both translations are bad. For each language, we collect 5 annotations per term. As shown in Table 10, `GPT-4o`'s candidate selection consistently outperforms majority voting across all five languages. These results highlight the necessity of `GPT-4o` in achieving high-quality translations for crowdsourced annotations.

However, human involvement remains indispensable. As discussed in Section 3.2, state-of-the-art LLMs and machine translation systems fail to provide consistent answers, rendering automatic translation through agreement across multiple models infeasible. Thus, by combining the expertise of LLMs as verifiers with humans as input sources, we achieve efficient, accurate, and reliable translations.

4.2 Task 2: Is Our Dataset Translation Better than 60-60?

The 60-60 initiative dataset also focuses on AI terminology translations and includes the five languages we consider. This overlap allows us to intersect the English terms in our dataset with those in the 60-60 evaluation dataset and compare their respective translations. We retrieve the translations from both datasets and conduct a comparative assessment through a crowdsourced evaluation, same as the process detailed in Section 4.1.

The results presented in Table 11 show that the translations in our dataset consistently and significantly outperform those from the 60-60 dataset across all five languages. Additionally, as shown in Table 12, annotators demonstrated fair agreement when rating the Arabic translations, and moderate agreement, with Fleiss' Kappa values ranging from 0.4 to 0.5, for translations in the other four languages. These findings underscore the superior quality of translations in our dataset compared

to the 60-60 dataset. Refer to Appendix B.4 for further details on the assessment of the two tasks.

5 Dataset Coverage Assessment

Domain Coverage We first investigate the AI domains covered by the collected terms in GIST and their distributions. We use `GPT-4o-mini` (Hurst et al., 2024) to identify the specific AI domains for each term, following the taxonomy proposed by Ding et al. (2023), which clusters the research domains of AI scholars. Figure 9 shows the distribution of the top six AI domains in GIST, with terms most frequently categorized into statistics, mathematics, computer science (CS), natural language processing (NLP), data science (DS), and computer vision (CV). We further embed all AI terminology in GIST using the `all-MiniLM-L6-v2` model (Reimers and Gurevych, 2019) and apply Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for dimensionality reduction to a two-dimensional space. The visualization in Figure 10 reveals that terms in domains such as NLP, CV, statistics, and mathematics form distinct clusters, while terms in CS and DS are more dispersed.

Terminology Coverage We surmise that our awarded paper dataset provides essential coverage of the terms in our terminology dictionary, where many terms are actually repeatedly extracted multiple times from different awarded papers. We present the rarefaction curve in Figure 11, which visually demonstrates the relationship between the number of extracted terms and the dataset size.

To statistically validate our claim, we conduct a one-sided one-sample t-test to compare the mean coverage ratio of the terminology dictionary. Specifically, we test whether the mean coverage ratio of randomly sampled subsets, constituting 60% of the terminology dictionary, is significantly above 80%. The null hypothesis (H_0) assumes that the mean coverage ratio of these subsets is less than or equal to 80%, while the alternative hypothesis (H_a) assumes that the mean coverage ratio exceeds 80%. Using 1,000 random samples, the analysis yields a t-statistic of 64.78 and a p-value of 0, rejecting the null hypothesis. These results provide statistical evidence supporting the sufficiency of our dataset in covering the terminology dictionary. Refer to Appendix B.5 for more details.

6 Experiments

We explore terminology integration approaches that do not need model retraining. We quantitatively evaluate two methods: terminology refinement via post-hoc LLM prompting and terminology substitution guided by word alignment.

6.1 Methods

Terminology Refinement: LLM Prompting With the recent advancements in multilingual and instruction-following capabilities of LLMs, it is now possible to leverage these models to refine translations. Specifically, we prompt the LLMs to revise the initial translation produced by a machine translation model, incorporating relevant term translations from our terminology dictionary within the provided context. For this refinement, we employ `GPT-4o-mini` and use the prompt illustrated in Figure 12. Refer to Appendix C.1 for more details.

Terminology Substitution: Word Alignment We apply the word alignment approach introduced by Dou and Neubig (2021) to identify and substitute term translations within the output. We use the multilingual BERT base model (Devlin et al., 2018) to tokenize both source and target sentences, and then process these inputs through its hidden layers to produce contextualized embeddings. We determine alignments by computing dot-product similarities between source and target token embeddings. High-confidence alignments are then filtered using a threshold of 1e-4, and subword alignments are aggregated to generate word-level mappings. Lastly, we directly replace identified term translations with those from GIST. We also conduct a post-hoc prompting after word alignment using `GPT-4o-mini` to render the translations morphologically coherent and correct. Refer to Appendix C.2 for more details. Additionally, we conduct a qualitative evaluation of traditional decoding techniques, including constrained beam search and token-level logits adjustment. See Appendix C.3 for more details.

6.2 Experiment Setup

Evaluation Set We conduct experiments on two evaluation sets: the **60-60 Evaluation Set** and the **AI Papers and Model Cards Evaluation Set**. The first evaluation set is from the 60-60 initiative, providing terminology translations into five target languages. As detailed in Section 4.2, our analysis

Model	Metric	Arabic			Chinese			French			Japanese			Russian		
		D	+P	+W	D	+P	+W	D	+P	+W	D	+P	+W	D	+P	+W
Evaluation Set: 60-60																
aya-expanse	BLEU	20.11	+ 1.23	+ 0.18	27.31	+ 1.33	+ 0.24	33.05	+ 2.46	+ 0.20	14.59	+ 0.61	+ 0.32	16.59	+ 1.59	- 0.05
	COMET	81.96	+ 0.71	- 0.52	83.43	+ 1.57	+ 0.08	81.83	+ 1.06	- 0.11	88.54	+ 0.32	- 0.01	82.27	+ 0.69	- 2.02
aya-23-8B	BLEU	19.98	+ 0.54	- 0.21	26.08	+ 0.47	+ 0.39	33.85	+ 2.28	- 0.11	15.06	+ 0.87	+ 0.36	15.77	+ 1.05	+ 0.37
	COMET	84.02	+ 0.81	- 0.24	85.12	+ 0.58	+ 0.38	82.40	+ 0.94	- 0.15	87.92	+ 0.50	+ 0.09	81.91	+ 0.40	- 2.26
gpt-4o-mini	BLEU	23.58	+ 1.07	- 0.00	32.64	+ 1.60	+ 0.66	40.80	+ 3.08	+ 0.50	21.46	+ 0.64	+ 0.19	17.25	+ 1.07	- 0.13
	COMET	85.77	+ 0.69	- 0.44	87.30	+ 0.48	+ 0.26	84.56	+ 0.68	- 0.04	89.96	+ 0.14	+ 0.01	83.68	+ 0.38	- 2.29
nllb	BLEU	22.38	+ 1.37	+ 0.64	17.29	+ 1.92	+ 1.02	34.93	+ 2.86	+ 0.21	6.19	+ 2.42	+ 0.53	17.30	+ 1.54	+ 1.07
	COMET	83.52	+ 0.83	- 0.45	78.22	+ 2.95	+ 0.73	82.83	+ 1.00	- 0.19	77.82	+ 3.80	+ 0.39	81.41	+ 0.97	- 1.54
seamless	BLEU	23.13	+ 1.16	- 0.03	26.26	+ 0.97	+ 0.80	40.04	+ 2.08	- 0.57	14.56	+ 0.74	+ 0.05	17.18	+ 1.71	+ 1.17
	COMET	84.07	+ 0.94	- 0.38	83.44	+ 1.48	+ 0.50	83.86	+ 0.78	- 0.07	85.05	+ 1.06	+ 0.16	82.33	+ 0.56	- 1.87
Evaluation Set: AI Papers & Model Cards																
aya-expanse	BLEU	11.47	+ 0.37	+ 0.10	12.04	+ 0.94	+ 0.17	18.84	+ 1.71	- 0.85	8.11	- 0.03	+ 0.04	13.84	+ 0.21	+ 0.32
	COMET	80.98	+ 0.46	- 0.51	82.42	+ 0.56	- 0.01	81.16	+ 0.36	- 0.79	85.48	+ 0.34	+ 0.07	82.76	+ 0.47	- 2.10
aya-23-8B	BLEU	14.28	+ 0.81	+ 0.28	14.50	+ 0.39	+ 0.19	24.49	+ 2.36	+ 0.08	9.22	+ 0.23	+ 0.35	16.39	+ 1.36	+ 0.61
	COMET	81.55	+ 1.03	- 0.62	83.88	+ 0.68	+ 0.04	82.55	+ 1.22	- 0.70	84.42	+ 0.81	+ 0.02	82.72	+ 1.14	- 2.08
gpt-4o-mini	BLEU	14.37	+ 0.53	- 0.42	17.21	+ 1.22	+ 1.39	24.45	+ 4.38	+ 1.28	10.55	+ 0.05	+ 0.03	18.02	+ 1.52	+ 0.40
	COMET	83.56	+ 0.86	- 0.19	86.08	+ 0.25	- 0.13	84.75	+ 0.33	- 1.03	87.91	+ 0.16	- 0.01	84.92	+ 0.44	- 2.02
nllb	BLEU	15.42	- 0.31	- 0.77	10.24	+ 2.19	+ 2.07	22.68	+ 2.73	+ 0.90	8.24	+ 1.08	+ 0.88	19.18	- 0.10	- 0.24
	COMET	81.23	+ 1.28	- 0.50	80.19	+ 1.61	+ 0.38	78.70	+ 3.81	- 1.59	83.05	+ 1.71	+ 0.70	80.46	+ 2.80	- 2.38
seamless	BLEU	15.38	+ 1.09	+ 0.45	13.67	+ 1.10	+ 0.73	24.34	+ 5.21	+ 1.49	9.42	+ 0.56	+ 0.42	18.43	+ 0.95	+ 0.35
	COMET	81.96	+ 1.18	- 0.39	80.70	+ 2.18	+ 0.16	83.76	+ 0.97	- 0.94	83.70	+ 0.88	+ 0.10	83.12	+ 1.50	- 1.79

Table 2: Evaluation results across five models and five languages using BLEU and COMET metrics. The first black value in each column represents the direct translation score (**D**). The second and third values, shown in red and green, indicate the relative performance change when applying the prompting-powered refinement method (**P**) and the word alignment method (**W**), respectively, compared to direct translation. See Table 15 for the complete results of additional metrics and ablations.

shows that our dataset achieves higher quality compared to the 60-60 set. Motivated by this, we enhance the labels of the 60-60 evaluation set by replacing AI terminology in their translations with `GPT-4o` while maintaining grammatical correctness and ensuring no loss in translation quality. This process generates updated ground truth labels for the evaluation.

The second evaluation set is manually created by combining text from two sources: 50 held-out AI research papers and 50 model cards generated by Liu et al. (2024). From this set, we randomly sampled 500 English text chunks from it for evaluation. To create ground truth labels for this set, we use Google Translate, a state-of-the-art machine translation model (Zhu et al., 2023; Santosa et al., 2024), to produce initial translations. We then prompt `GPT-4o` to refine these translations by updating AI terminology. This evaluation set is designed to explore the application of our AI terminology dictionary in two major domains: AI research papers and model cards. Both evaluation set references have been verified by human experts and exhibit high quality. Further details are provided in Appendix C.4.

Models We evaluate the following models: `gpt-4o-mini` (Hurst et al., 2024), `hf-seamless-m4t-large` (Barrault et al., 2023), `nllb-200-3.3B` (Costa-jussà et al., 2022), `aya-23-8B` (Aryabumi et al., 2024), and `aya-expanse-8B` (Dang et al., 2024). While the prompting method applies to all models, the word alignment method is not applicable to `gpt-4o-mini`, as it does not provide access to its model weights.

Evaluation Metrics We adopt BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), ChrF (Popović, 2015), ChrF++ (Popović, 2017), and TER metrics (Snover et al., 2006) for the quantitative evaluation of translation quality, following the methodology of Salesky et al. (2023). We utilize the `wmt22-comet-da` model (Rei et al., 2022) to compute COMET scores.

6.3 Experiment Results

Table 2 presents the quantitative evaluation results under the experimental settings described earlier. We draw several key observations:

First, the results across the two evaluation sets show consistency, indicating the robustness of the findings. Among the models, `gpt-4o-mini`

Figure 3: The interface demonstrates a paper example from Tang et al. (2024) and introduces a new translation feature on the ACL Anthology website. Non-English speakers can click the "Translate" button at the bottom right of the webpage and select their preferred language for translation. The page dynamically displays two translations side by side: one using direct translation from a machine translation model (e.g., `seamless` in the figure) and the other enhanced with prompting-powered refinement applied to the left translation. Highlighted text indicates updated terminology integrated from GIST, showcasing improved translation performance.

achieves the best overall translation performance. Among the remaining four models, differences in performance are less pronounced.

Several trends emerge when comparing different terminology integration approaches. The prompting-powered refinement method consistently outperforms direct translation across nearly all languages, models, and evaluation metrics, which highlights its effectiveness in incorporating AI-specific terminology into translations.

The word alignment method demonstrates mixed performance, showing improvements for Chinese and Japanese translations but leading to declines for Arabic, French, and Russian. This discrepancy is due to linguistic differences: Chinese exhibits minimal morphological changes, allowing straightforward substitution of terminology with limited disruption to surrounding syntax; In contrast, languages like Arabic often require agreement in gender, number, and syntactic roles, making noun replacement more complex and error-prone. These findings underscore the superior performance of the prompt-powered refinement method, as well as the importance of tailoring terminology integration approaches to the linguistic characteristics of target languages. See Appendices D.1 and D.2 for further discussion of the results. We further validate our findings using a one-sided paired t-test in Appendix D.3.

7 Website Demonstration

To facilitate real-world usage of our AI terminology dictionary, we built on the 60-60 initiative by modifying the ACL Anthology website layout and introducing a new terminology translation feature,[6] as illustrated in Figure 3. The ACL Anthology website was chosen for this demonstration due to its extensive collection of AI-related research papers, making it an ideal platform to showcase the potential impact of our work. In this demonstration, translations of key terms are refined and standardized based on the terminology dictionary we developed, providing more accurate terminology translations.This enhancement represents a step toward improving access to AI knowledge for non-English speakers by offering a consistent and reliable translation system, and broadens the accessibility of AI research to a global audience.

8 Conclusion

We present GIST, a large-scale multilingual AI terminology dataset addressing gaps in translating AI-specific terms. Combining LLM-based extraction and validation with human expertise, it includes 5K English terms with translations into five languages, surpassing the ACL 60-60 benchmark. LLM prompting proved effective for post-hoc terminology integration, improving translation

[6] https://acl6060.org/

quality across five metrics. We also provide a website demonstration to enhance the accessibility of our work for non-English speakers, supporting equitable AI knowledge access and fostering global collaboration in AI research.

Limitations

This work is subject to several limitations. First, our dataset assumes a one-to-one correspondence between English terms and their translations, which does not account for cases where multiple equally valid translations exist for a single term. This simplification may overlook the nuanced variations in AI terminology usage across languages.

Despite significant efforts in collecting terminologies, our coverage is not exhaustive. The field of artificial intelligence lacks a well-defined boundary, making it challenging to ensure comprehensive inclusion of all relevant AI terms. Additionally, while we provide translations for five widely used languages, this represents only a subset of the global linguistic diversity and leaves many other languages unaddressed.

Furthermore, while our methodology is tailored for AI terminology translation, its application to other domains may require adaptation. Although our LLM + Human hybrid framework for data collection, translation, and evaluation is broadly applicable and does not rely on AI-specific models or tools, domain-specific terminology translation poses unique challenges, such as variations in terminology collection and the need for domain-specific expertise in evaluation.

Nonetheless, we hope this work inspires the community to further advance the creation and refinement of multilingual AI terminology dictionaries, addressing these limitations and extending coverage to more languages and domains.

Ethical Considerations

We manually collected awarded papers from official conference websites. Only papers available under open-source licenses for research use were downloaded. Similarly, model card contents used in this study were sourced from openly shared materials by their respective authors.

During data processing, we ensured that no personal information, such as human names irrelevant to AI methods or metrics, was included in the dataset. Automatic and manual reviews were conducted to verify the exclusion of any sensitive or private details.

For term translation, we employed LLMs to assist with extraction and validation. While we acknowledge that LLMs may exhibit biases when selecting the best translations, manual evaluations confirmed the superior performance of LLMs in this task.

In crowdsourced experiments, we respected participant privacy by not collecting any demographic information. All contributors were fairly compensated according to MTurk's payment standards. Additionally, our collected terminology dataset does not involve ethically sensitive or controversial content, focusing exclusively on technical terms relevant to AI.

Acknowledgments

We extend our gratitude to the Minister of State for Artificial Intelligence of the United Arab Emirates for making the Arabic AI terminology dictionary publicly available. We thank DataFranca for granting us access to their French AI terminology dictionary. We thank Jiqizhixin for sharing their Chinese AI terminology database. We also thank the Russian Association of Artificial Intelligence for providing access to the Russian AI terminology dictionary. For more details, please refer to Appendix B.2.

References

Andreas Abel and Christian M. Meyer. 2013. The dynamics outside the paper: user contributions to online dictionaries.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Anwar AH Al-Athwary and Husam Khaled Ali. 2024. Arabicization via loan translation: a corpus-based analysis of neologisms translated from english into arabic in the field of information technology. *Open Cultural Studies*, 8(1):20240005.

Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, et al. 2023. The manifold costs of being a non-native english speaker in science. *PLoS Biology*, 21(7):e3002184.

Ulrich Ammon. 2003. Global english and the non-native speaker. *A: tonKin, Humphrey ir eaGan, Timo-*

thy (eds.). Language in the 21st century. Amsterdam: John Benjamins, pàg, pages 23–34.

Ulrich Ammon. 2012. Linguistic inequality and its effects on participation in scientific discourse and on global knowledge accumulation – with a closer look at the problems of the second-rank language communities. *Applied Linguistics Review*, 3(2):333–355.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*.

Anthropic. 2024. Introducing the next generation of claude.

Mihael Arcan, Claudio Giuliano, Marco Turchi, and Paul Buitelaar. 2014. Identification of bilingual terms from monolingual documents for statistical machine translation. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*, pages 22–31, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Awadh Nasser Munassar Awadh. 2024. Challenges and strategies of translating scientific texts: A comparative study of human translation and artificial intelligence. *Educational Administration: Theory and Practice*, 30(4):9898–9909.

Anees Bahji, Laura Acion, Anne-Marie Laslett, and Bryon Adinoff. 2023. Exclusion of the non-english-speaking world from the scientific literature: Recommendations for change for addiction journals and publishers. *Nordic Studies on Alcohol and Drugs*, 40(1):6–13.

Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya, and Siddhartha Bhattacharyya. 2019. Assembling translations from multi-engine machine translation outputs. *Applied Soft Computing*, 78:230–239.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Andrea Calabrese and W Leo Wetzels. 2009. Loan phonology: Issues and controversies. In *Loan phonology*, pages 1–10. John Benjamins Publishing Company.

Husanjot Chahal, Jennifer Melot, Sara Abdulla, Zachary Arnold, and Ilya Rahkovsky. 2022. Country activity tracker.

Elsie KY Chan, John SY Lee, Chester Cheng, and Benjamin K Tsou. 2023. Post-editing of technical terms based on bilingual example sentences. In *Machine Translation Summit XIX (MT Summit 2023)*, pages 385–392.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the second conference on machine translation*, pages 157–168.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. *arXiv preprint arXiv:2004.14837*.

Ksenia Chmutina, Neil Sadler, Jason von Meding, and Amer Hamad Issa Abukhalaf. 2021. Lost (and found?) in translation: key terminology in disaster studies. *Disaster Prevention and Management: An International Journal*, 30(2):149–162.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. *arXiv preprint arXiv:2410.14057*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ido Dagan and Kenneth Church. 1994. Termight: Identifying and translating technical terminology. In *Fourth Conference on Applied Natural Language Processing*, pages 34–40.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yiwen Ding, Jiarui Liu, Zhiheng Lyu, Kun Zhang, Bernhard Schoelkopf, Zhijing Jin, and Rada Mihalcea. 2023. Voices of her: Analyzing gender differences in the ai publication world. *arXiv preprint arXiv:2305.14597*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. *arXiv preprint arXiv:1906.01105*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Zhaopeng Feng, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. 2024. Ladder: A model-agnostic framework boosting LLM-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393, Miami, Florida, USA. Association for Computational Linguistics.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*.

Javier Ferrando, Gerard I Gállego, Belen Alastruey, Carlos Escolano, and Marta R Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. *arXiv preprint arXiv:2205.11631*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Julie Giguere. 2023. Leveraging large language models to extract terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 57–60.

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.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way. 2020a. Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation*, 34:149–195.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020b. Terminology-aware sentence mining for nmt domain adaptation: Adapt's submission to the adapmt 2020 english-to-hindi ai translation shared task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23.

Rejwanul Haque, Sergio Penkale, and Andy Way. 2018. Termfinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52(2):365–400.

Eva Hasler, Adrià De Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. *arXiv preprint arXiv:1805.03750*.

Chris Hokamp and Qun Liu. 2017a. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Chris Hokamp and Qun Liu. 2017b. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yiping Jin, Min-Yen Kan, Jun-Ping Ng, and Xiangnan He. 2013. Mining scientific terms and their definitions: A study of the ACL Anthology. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 780–790, Seattle, Washington, USA. Association for Computational Linguistics.

Abdullah A Khuwaileh and Tariq Khwaileh. 2011. It terminology, translation, and semiotic levels: Cultural, lexicographic, and linguistic problems.

JiWoo Kim, Yunsu Kim, and JinYeong Bak. 2024. KpopMT: Translation dataset with terminology for kpop fandom. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 37–43, Bangkok, Thailand. Association for Computational Linguistics.

Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024. Automatic generation of model and data cards: A step towards responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1975–1997, Mexico City, Mexico. Association for Computational Linguistics.

Enrico Manzini, Jon Garrido-Aguirre, Jordi Fonollosa, and Alexandre Perera-Lluna. 2022. Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. *Expert Systems with Applications*, 204:117446.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911.

Palanichamy Naveen and Pavel Trojovskỳ. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

S Nishio, H Nonaka, N Tsuchiya, A Migita, Y Banno, T Hayashi, H Sakaji, T Sakumoto, and K Watabe. 2024. Extraction of research objectives, machine learning model names, and dataset names from academic papers and analysis of their interrelationships using llm and network analysis. *arXiv preprint arXiv:2408.12097*.

OpenAI. 2023. New models and developer products announced at devday.

OpenAI. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *arXiv preprint arXiv:1804.06609*.

Fernando Prieto Ramos. 2021. Translating legal terminology and phraseology: between inter-systemic incongruity and multilingual harmonization.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Made Hery Santosa, Gusti Ayu Made Trisna Yanti, and Luh Diah Surya Adnyani. 2024. The integration of google translate as a machine translation aid in efl students'thesis composition. *LLT Journal: A Journal on Language and Language Teaching*, 27(1):214–229.

Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2024. TMU-HIT's submission for the WMT24 quality estimation shared

task: Is GPT-4 a good evaluator for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pages 529–534, Miami, Florida, USA. Association for Computational Linguistics.

Anne-Kathrin Schumann and Héctor Martínez Alonso. 2018. Automatic annotation of semantic term types in the complete ACL Anthology reference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mahsa Shamsabadi, Jennifer D'Souza, and Sören Auer. 2024. Large language models for scientific information extraction: An empirical study for virology. *arXiv preprint arXiv:2401.10040*.

Mark Shuttleworth. 2014. *Dictionary of translation studies*. Routledge.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742, Mexico City, Mexico. Association for Computational Linguistics.

Irsha Tehseen, Ghulam Rasool Tahir, Khadija Shakeel, and Mubbashir Ali. 2018. Corpus based machine translation for scientific text. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 196–206. Springer.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47:9–31.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Richard Yue, John Ortega, and Kenneth Church. 2024. On translating technical terminology: A translation workflow for machine-translated acronyms. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 48–54, Chicago, USA. Association for Machine Translation in the Americas.

Omar Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1220–1229.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2019. Adding interpretable attention to neural translation models improves word alignment. *arXiv preprint arXiv:1901.11359*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How good are llms for literary translation, really? literary translation evaluation with humans and llms. *arXiv preprint arXiv:2410.18697*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Further Discussion on AI Terminology Translation

In many non-English texts nowadays, AI terminology is often borrowed as untranslated loanwords. However, translating these terms into local languages is crucial for the following reasons:

1. Translation democratizes AI for non-English speakers. Translating AI terminology into native languages plays a vital role in making the field more accessible to non-English-speaking communities. For those outside the AI field, translating these terms fosters a clearer understanding, as it is easier for them to grasp complex concepts in their own language rather than through borrowed English terms. Prior studies highlight that translating English loanwords into Arabic and other languages helps to fill lexical gaps (Calabrese and

Wetzels, 2009; Al-Athwary and Ali, 2024), making terms more meaningful while preserving linguistic and cultural identities. This practice promotes the standardization of AI terminology, reduces dependency on foreign languages, and improves educational accessibility for non-English speakers.

2. Our work aligns with and extends the ACL 60-60 initiative (Salesky et al., 2023), which aims to translate AI-specific terms to enhance multilingual scientific communication and foster AI inclusivity. By expanding this effort, we seek to improve the scale, scope, and quality of AI terminology translation, furthering the global reach of AI research.

3. The prevalence of untranslated loanwords varies significantly across languages. The prevalence of untranslated loanwords is more common in languages influenced by English or with linguistic similarities to English. However, this is not the case for many other languages. As a concession to this variability, we employed `GPT-4o-mini` to determine whether terms in the GIST dataset should be translated. Table 3 presents the results, based on the prompt in Figure 4, which highlights the disparities across languages.

Language	Percentage of Terms Requiring Translation
Chinese	95.0%
Arabic	73.0%
French	72.0%
Japanese	53.8%
Russian	80.5%

Table 3: Proportion of AI terms requiring translation across different languages.

B Additional Dataset Details

B.1 Additional Dataset Statistics

To better understand the composition and characteristics of the dataset, we performed a comprehensive statistical analysis across multiple dimensions, including domain distribution, semantic clustering, and lexical structure across five languages: Arabic, Chinese, French, Japanese, and Russian. Below, we present key findings through detailed visualizations and lexical statistics.

Lexical Statistics. Table 4 summarizes the lexical characteristics of the dataset across the five languages. Key metrics include:

- **Number of Terms:** The dataset contains 4,844 to 6,527 terms per language, with French and Chinese having the largest repositories.

- **Unique Words:** English terms comprise 2,400–3,400 unique words across datasets. Target languages exhibit varying lexical diversity, with French and Russian showing higher uniqueness due to linguistic richness.

- **Words per Term:** On average, English terms consist of approximately two words (~ 2.02), while target languages show higher variability. French terms, for instance, require more words (2.68 ± 1.19), reflecting language-specific expansion during translation.

- **Characters per Term:** English terms maintain consistent lengths ($\sim$ 17 characters), while target languages vary significantly. For example, Chinese terms are concise (4.66 ± 1.96) due to its logographic script, whereas French (21.27 ± 8.49) and Russian (20.20 ± 7.83) terms are longer, reflecting the morphology of these languages.

To tokenize terms, we utilized `nltk.word_tokenize` for English, Arabic, French, and Russian; `jieba` for Chinese; and `MeCab Owakati` for Japanese. These tools ensured language-specific tokenization accuracy, enabling detailed lexical analysis.

Terminology Examples To identify and show the most frequently used terms in the original set of awarded papers, we extracted and ranked the top 150 terminologies based on their occurrence frequency. Table 5 provides a comprehensive list of these terms, categorized by their rank and grouped for clarity.

Temporal Statistics We analyze the distribution of terms in GIST based on the publication years of the papers. Table 6 presents the top 10 most frequent terms in awarded papers for each year. This analysis offers insights into the temporal evolution of AI terminology.

You are an expert in {tgt_lang} AI scientific literature. Determine whether the term "{term}" is more commonly:

A: Translated into {tgt_lang}.
B: Borrowed from English as a loanword.

Choose B only if none of the words in the term are translated and the entire term is used as-is in English. Output only "A" or "B", based on what is most prevalent in {tgt_lang} AI academic and technical contexts.

Figure 4: Prompt used to determine whether GIST terms should be translated.

	Arabic	Chinese	French	Japanese	Russian
# Terms	4844	6426	6527	4770	5167
Unique En Words	2470	3244	3470	2424	2615
Unique Tgt Words	3161	2838	4036	2050	4210
En Words/Term	2.02 ± 0.59	2.05 ± 0.68	2.07 ± 0.67	2.02 ± 0.58	2.01 ± 0.59
Tgt Words/Term	2.36 ± 0.83	2.26 ± 0.90	2.68 ± 1.19	2.53 ± 0.98	2.16 ± 0.80
En Chars/Term	16.99 ± 5.97	17.26 ± 6.60	17.44 ± 6.57	16.96 ± 5.91	16.94 ± 5.90
Tgt Chars/Term	15.22 ± 5.66	4.66 ± 1.96	21.27 ± 8.49	6.89 ± 3.16	20.20 ± 7.83

Table 4: Lexical statistics of the dataset across languages, including standard deviations. "En" denotes English, and "Tgt" denotes the target language. Terms are tokenized into words using `nltk.word_tokenize` for English, Arabic, French, and Russian, `jieba` for Chinese, and the `MeCab Owakati` tokenizer for Japanese.

Conclusion. The statistical analysis highlights the diversity and interdisciplinary nature of the dataset. Figures 9 and 10 illustrate domain-wise distributions and semantic clusters, while Table 4 quantifies lexical variations across languages. Together, these findings provide a robust understanding of the dataset's structure, supporting its utility for multilingual and domain-specific AI applications.

B.2 Terminology Collection Details

We selected terms from awarded papers as this approach provides an efficient and manageable way to curate a representative sample of influential AI research. While we acknowledge that valuable AI terminology also exists in non-awarded yet influential papers, capturing all relevant terms across the vast AI field is infeasible. Nonetheless, we believe our dataset offers a comprehensive representation in terms of domain coverage and unique terminology. We considered expanding our selection to include highly cited papers; however, to our knowledge, no automated method reliably identifies such papers. Even if one existed, determining an appropriate citation threshold for inclusion would remain a challenge.

To refine our terminology selection, we employ a two-step procedure to filter out non-nominal phrases. First, we prompt `GPT-4` to retain only nouns or noun phrases as candidates. Subsequently, we use a POS tagger to further remove any phrases that do not contain a noun.

We integrate terms from the WikiPedia Glossary of AI[7], which serves as a comprehensive starting point for artificial intelligence (AI)-related terminology, ensuring alignment with globally recognized concepts and definitions. To enhance multilingual coverage and domain relevance, we also include terms from several other specialized AI terminology initiatives across different languages:

Arabic AI Dictionary[8]: Published by the Arabic Government AI Office, this resource aims to elevate the status of the Arabic language in AI, standardize terminology, reduce linguistic ambiguity, and foster better integration of Arabic speakers into the global AI community. By providing accurate translations and clear definitions for English AI terms, this dictionary promotes knowledge dissemination and encourages collaboration within the Arabic-speaking AI ecosystem.

Chinese GitBook AI Term Database[9]: Developed by Jiqizhixin (Machine Heart), this database represents an extensive effort to document technical terms encountered during the translation of AI

[7] en.wikipedia.org/wiki/Glossary_of_artificial_intelligence
[8] ai.gov.ae/ar/ai-dictionary/
[9] https://jiqizhixin.gitbook.io/artificial-intelligence-terminology-database

articles and research papers. Starting with practical usage, the project has evolved to incorporate domain-specific expansions based on authoritative textbooks and expert input, offering the Chinese AI community a unified and precise reference for both academic and industrial applications.

French AI Dictionary[10]**:** The first comprehensive French reference tool for data science and AI, this dictionary addresses the needs of public service, commerce, research, and education. It aims to bridge the gap between French and English AI terminologies, ensuring accessibility and standardization for French-speaking professionals and researchers.

Russian AI Dictionary[11]**:** This initiative captures the interdisciplinary nature of AI by including terminology drawn from fields such as logic, psychology, linguistics, and cybernetics. Leveraging contributions from Russian and Soviet experts, this dictionary emphasizes the frequency and relevance of terms within AI-specific contexts, providing a culturally adapted yet globally aligned resource for Russian-speaking researchers.

B.3 Terminology Translation Details

Terminology Dataset Collection In this section, we provide a detailed description of our methodology for creating the multilingual AI terminology dictionary and the associated translation experiments. The terms were extracted from papers published in the top AI conferences as shown in Table 7 across various fields including Artificial Intelligence, Computer Vision, Machine Learning, Natural Language Processing, and Web & Information Retrieval. These conferences represent the leading venues in their respective domains.

Prompt Design for Translation To ensure consistent and high-quality translations, we used carefully designed prompts for GPT-3.5-Turbo and Claude 3 Sonnet. We provided sentence contexts relevant to each AI terminology, split into a maximum of three chunks, each containing up to 64 words. As shown in Figure 6, the translation prompt asks the model to translate an AI-specific term into a target language, with context provided to clarify the meaning. If the term is an abbreviation or a technical term that should remain in its original form to avoid confusion, the models are instructed to retain the English term.

Translation Agreement Analysis We evaluated the consistency of translations across three models: Claude 3 Sonnet, GPT-3.5-Turbo, and Google Translate API. Table 8 summarizes the agreement ratios for five target languages: Arabic, Chinese, French, Japanese, and Russian. The results indicate significant variation in agreement ratios across languages, with Chinese achieving the highest three-model agreement (42.71%) and Arabic having the lowest (10.11%). These findings underscore the need for human involvement in AI terminology translation, as automatic translations often fall short due to the inherent limitations of current models.

To further assess translation quality, we also examined which model's outputs align most closely with human annotations by incorporating `GPT-4o`. Table 9 presents the results, showing that `GPT-4o` is the most advanced LLM available at the time of this study. This claim is further supported by relevant literature (Wang et al., 2024; Sato et al., 2024; Zhang et al., 2024; OpenAI, 2024). However, our evaluation reinforces the argument that, despite GPT-4o's advancements, it remains inadequate for precise terminology translation. This highlights the necessity of human annotations to ensure translation accuracy and reliability.

Human Translation via Mturk To further validate the translations, we employed human translator through the MTurk platform. Each terminology was translated by 10 independent annotators, resulting in a set of 10 translation candidates for each term. The platform layout and guidance for helping translators give these translations is shown in Figure 7. We compensate annotators in accordance with MTurk's payment standards.

B.4 Dataset Quality Assessment Details

To evaluate the quality of translations in our dataset, we conducted two tasks involving human annotators. Annotators were presented with pairs of translations generated by different methods and were tasked with evaluating their relative quality. The analysis of their ratings is summarized in Tables 10, 11, and 12. Below, we describe the findings in detail.

Task 1: Comparison Between GPT-4o-Selected Candidates and Majority-Voted Translations In Task 1, we constructed the evaluation dataset by

[10] https://datafranca.org/wiki/Cat%C3%A9gorie:GRAND_LEXIQUE_FRAN%C3%87AIS

[11] https://www.raai.org/pages/UGFnZVR5cGU6MTAwMw==

Your task is to identify AI scientific terminology from the paragraph below, based on the provided definition:
Here is the definition of AI scientific terminology:
```
AI scientific terminology refers to specialized nouns or noun phrases within Artificial Intelligence, encompassing essential concepts, methods, models, algorithms, and systems. These terms must:
1. Be composed of nouns, adjectives, and occasionally prepositions.
2. Be context-specific to AI, having either no meaning or a different meaning outside this field. Additionally, these terms often pose significant challenges for accurate translation by machine translation models due to their technical specificity.
```
Provide only the identified terms in your response, separated by commas. If no scientific terms are found, respond with "None" only. Example: transformer, batch normalization, embedding.

Here is the paragraph:
```
{text}
```

Figure 5: Prompts for extracting AI terminologies with `LLaMA-3-70B-Instruct`.

Translate the following AI scientific term directly into {tgt_lang} based on its context. If the term is an abbreviation that could reduce confusion by keeping it as is, provide the original English term. Directly provide your translated term without any explanations.

- Term: {term}
- Context: "{context}"

Figure 6: Prompts for translating AI terminologies with `GPT-3.5-Turbo` and `Claude 3 Sonnet`.

first randomly sampling 1,000 terms per language that had received majority votes. We then filtered out terms where both translation strategies produced identical results, leaving approximately 200 terms per language for human evaluation. Specifically, this resulted in 323 Arabic terms, 185 Chinese terms, 180 French terms, 206 Japanese terms, and 230 Russian terms.

Annotators compared translations selected by GPT-4o (Method 1) with majority-voted translations generated by human annotators (Method 2). The choices were: A. Both translations are good; B. Method 1's translation is better; C. Method 2's translation is better; D. Both translations are bad.

Table 10 summarizes the distribution of annotators' choices for Task 1. Across all languages, the majority of annotations fell into category A, where both translations were rated as good. Japanese showed the highest percentage of agreement in this category (56.99%), followed by Russian (54.43%) and Chinese (50.59%). Cases where Method 1 translations were rated as better (category B) ranged from 24.37% (Japanese) to 30.26% (Russian). Similarly, cases where Method 2 translations were rated as better (category C) ranged from 13.04% (Russian) to 20.37% (Arabic). The lowest percentage of responses was observed in category D, where both translations were rated as bad, accounting for less than 5% of responses across all languages.

Task 2: Comparison Between Dataset Translations and 60-60 Initiative Translations In Task 2, the number of overlapping terms between the 60-60 evaluation set and our dataset varies across languages for two main reasons. First, some terminology translations appear in certain languages but not others due to the integration of external data sources, as shown in Table 1. Second, some translations are identical in both our dataset and the 60-60 evaluation set; these were removed prior to human evaluation. As a result, the final analysis includes 162 Arabic terms, 106 Chinese terms, 77

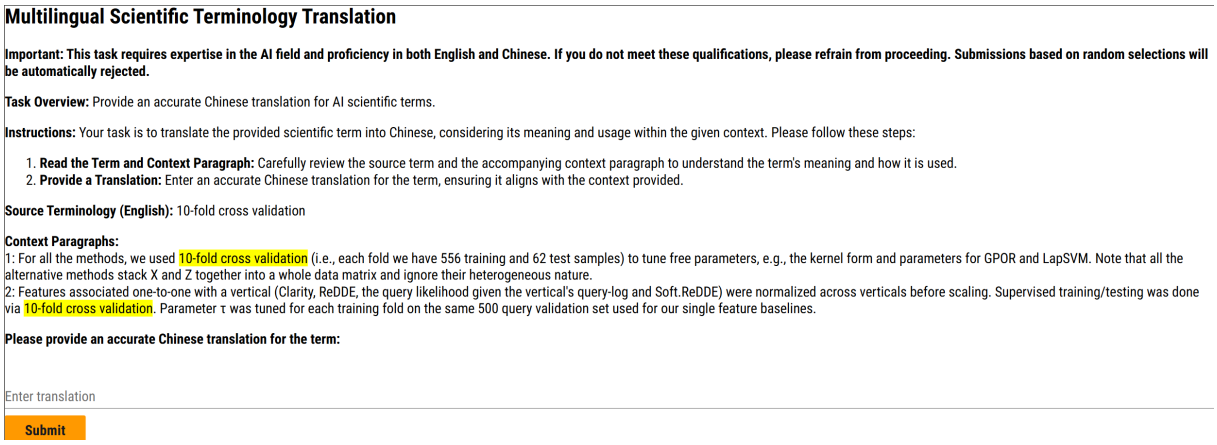

Figure 7: The MTurk layout demonstration for AI terminology translation generation task.

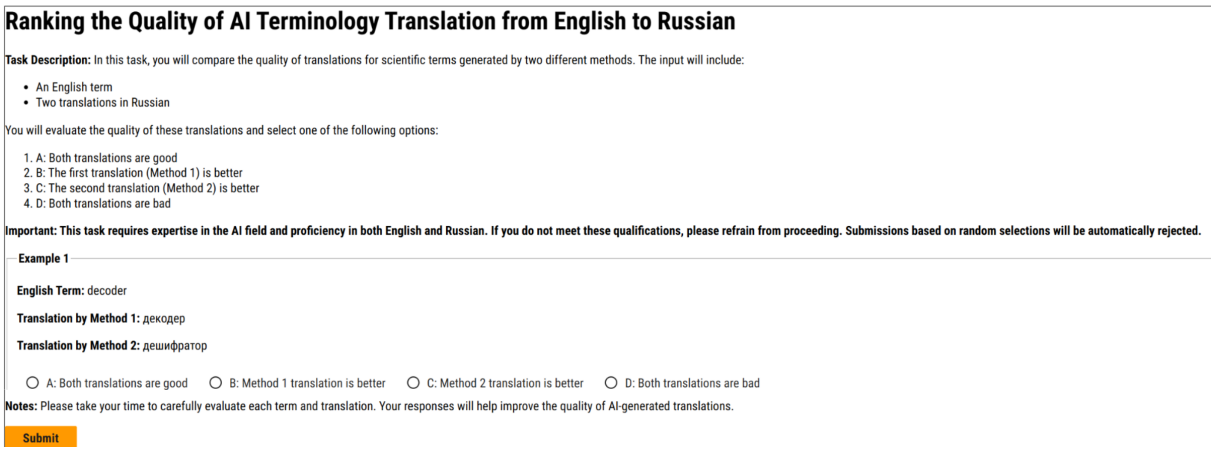

Figure 8: The MTurk layout demonstration for evaluating AI terminology translations. The layout is used for two tasks: (1) comparing the GPT-4o-selected candidate with the majority-voted candidate, and (2) comparing the translations in our dataset with those in the 60-60 initiative evaluation set.

French terms, 103 Japanese terms, and 88 Russian terms.

Annotators compared translations from our dataset (Method 1) with those in the 60-60 initiative evaluation set (Method 2). The distribution of choices is shown in Table 11. In this task, a higher percentage of annotators preferred translations from Method 2 (category B) for most languages, especially in Chinese (43.02%), French (43.64%), and Russian (45.00%). However, Japanese translations from Method 1 had a significantly higher percentage in category A, with 57.28% of annotators agreeing that both translations were good. Instances where both translations were rated as bad (category D) remained low across all languages, ranging from 2.60% to 5.68%.

Inter-Annotator Agreement To measure inter-annotator agreement, we calculated Fleiss' Kappa scores for each language in both tasks. Table 12 reports these values. Fleiss' Kappa values between $0.20 \leq \kappa < 0.40$ indicate fair agreement, while values between $0.40 \leq \kappa < 0.60$ indicate moderate agreement. In Task 1, Kappa scores ranged from 0.21 (Japanese) to 0.39 (Russian), showing fair agreement across languages. In Task 2, Kappa scores improved, ranging from 0.39 (Japanese) to 0.50 (French), indicating moderate agreement for most languages.

Conclusion The results demonstrate that our dataset's translations are of high quality, with a majority of annotators rating them as good. While inter-annotator agreement was fair in Task 1, moderate agreement was observed in Task 2, highlighting the robustness of our dataset compared to established benchmarks.

B.5 Dataset Coverage Assessment Details

Domain Distribution. The terminology in the dataset spans various AI-related domains. As shown in Figure 9, the six most frequent domains include Statistics and Probability (13.31%), Math (12.24%), Computer Science (11.74%), Nat-

ural Language Processing (11.50%), Data Science (9.98%), and Computer Vision (6.57%). The largest proportion of terms (34.65%) falls under the "Other" category, which represents interdisciplinary or less-defined concepts that do not fit neatly into any of the predefined categories. This distribution reflects the diversity and multidisciplinary nature of GIST.

Semantic Clustering. We used Uniform Manifold Approximation and Projection (UMAP) to visualize the semantic relationships between terms across domains. Figure 10 shows a low-dimensional embedding of terms, where each point represents a term, color-coded by its domain. The visualization reveals distinct clusters corresponding to each domain, indicating strong intra-domain coherence. Overlaps between clusters (e.g., Data Science and Natural Language Processing) highlight the interconnected nature of these fields, where concepts are often shared or applied across domains.

Coverage Analysis through Rarefaction. To assess the comprehensiveness of our terminology collection, we conducted a rarefaction analysis, which is visualized in Figure 11. This analysis reveals the relationship between the sample size of papers and the coverage of AI terminology. The curve demonstrates a characteristic asymptotic behavior, starting with a steep increase in coverage ratio from approximately 0.3 at 10% of papers to 0.6 at 30% of papers. As the subset size increases, the marginal gain in coverage gradually diminishes, reaching a coverage ratio of approximately 0.9 at 100% of the papers. The error bars, representing standard deviation across 50 random samples for each subset size, notably decrease as the sample size increases, indicating more stable coverage at larger sample sizes. The asymptotic nature of the curve approaching 0.9 coverage suggests that our current collection has achieved a robust representation of commonly used AI terminology, with additional papers likely to introduce increasingly specialized or niche terms at a decreasing rate. The relatively small error bars at larger sample sizes (80-100%) indicate high consistency in terminology coverage across different subsets of the literature, supporting the reliability of our collection methodology.

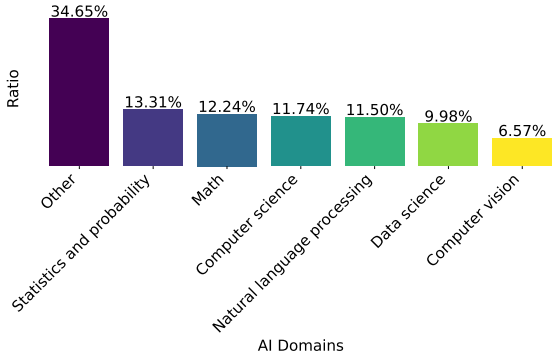

Figure 9: Terminology distribution of the top-6 AI domains in GIST.

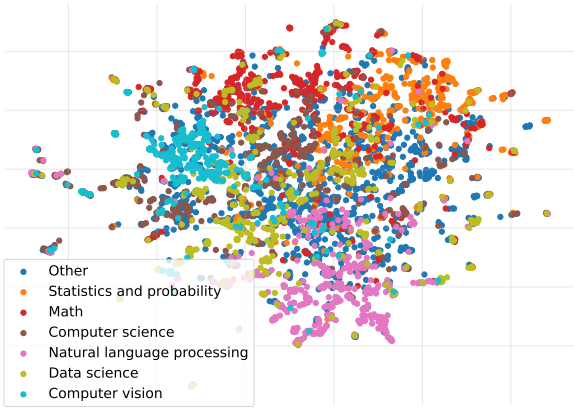

Figure 10: UMap visualization of terms in GIST by domain.

C Additional Experiment Details

C.1 Terminology Refinement via Prompting Details

In this subsection, we describe the process of refining machine translation outputs by leveraging prompts tailored for terminology consistency. The objective is to ensure that specific domain-related terms are translated accurately, adhering to predefined mappings provided in the dataset.

As illustrated in Figure 12, the prompt incorporates a term dictionary, the source language `src_lang` text, and the initial machine-translated `tgt_lang` output. The model is instructed to revise the translation by applying the specified target terms for corresponding source terms while maintaining the rest of the content unchanged. This approach ensures consistency and accuracy in translation outputs, particularly in specialized fields, by integrating domain knowledge directly into the model's refinement process.

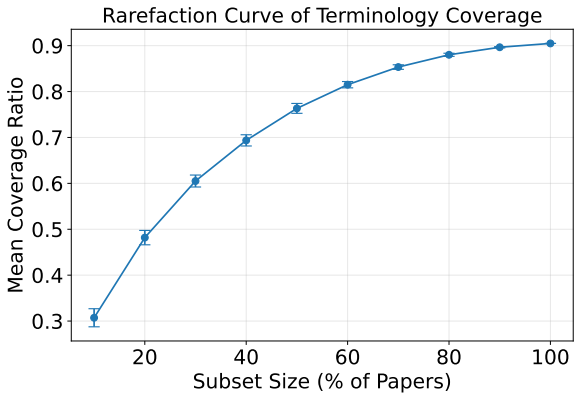

Figure 11: Rarefaction curve showing the mean coverage ratio of AI terminology with respect to the subset size of papers. Each subset size is sampled 50 times, and the error bars represent the standard deviation of the coverage ratios across these samples.

C.2 Terminology Substitution by Word Alignment Details

To validate our analysis in Section 6.3 regarding the varying performance of the word alignment method across different languages, we perform a post-hoc prompting step after word alignment. This step ensures that the translations are morphologically coherent and accurate, and is used to compare with the original results. The prompt is depicted in Figure 13. Refer to Appendix D.1 for complete experiment results.

C.3 Terminology-Aware Decoding Approaches

For constrained beam search (Hokamp and Liu, 2017b), we use the Hugging Face implementation, which enforces the inclusion of term translations from GIST in the output.[12] For token-level adjustment, we first identify AI terminology and their corresponding expected translations from GIST. We then modify the output logits of these tokens by increasing them by factors of 10/7, 10/8, and 10/9 to prioritize their selection during decoding. See Appendix D.2 for a discussion on performance.

C.4 Evaluation Set Creation Details

The AI Papers and Model Cards evaluation set is created using 50 held-out AI research papers and 50 model cards generated by Liu et al. (2024). We then segment the content into sentence chunks, ensuring that each chunk contains no more than 64 words. From this set, we randomly sample 500 sentence chunks for evaluation.

For both evaluation sets, we conducted a manual evaluation, involving a new group of five expert annotators on MTurk. We randomly sampled 50 examples from the 60-60 test set and another 50 examples from the AI papers + model card test set. The annotators were provided with the English sentences, their corresponding translations, the associated terminologies, and their translations. The evaluation included the following tasks:

- Task 1: Overall Translation Quality (rated on a scale from 1 to 5)
  - 1 = The translation makes no sense
  - 2 = Poor translation quality
  - 3 = Acceptable translation quality
  - 4 = Good translation quality
  - 5 = Excellent translation quality
- Task 2: Grammatical and Morphological Correctness (binary annotation: Yes/No)
- Task 3: Accuracy of Terminology Translations (binary annotation: Yes/No)

The results of the manual evaluation are summarized in Table 13. We observe consistently high performance across all languages in terms of translation quality, grammatical correctness, and terminology accuracy, highlighting the robustness and linguistic reliability of ground truth labels generated in our evaluation set. We don't measure the impact of incorporating GIST because there are only two possible reasons why translation quality might decrease after using GPT-4o to integrate terminologies: (1) low-quality terminology translations in our glossary, and (2) GPT-4o failing to integrate terminologies while maintaining grammatical accuracy. Regarding (1), we have already validated our glossary's translation quality in Section 4. Regarding (2), terminology integration is a straightforward task, and it is unlikely that GPT-4o would significantly degrade grammatical accuracy.

D Additional Experiment Results

D.1 Additional Quantitative Results

We present comprehensive evaluation results using ChrF (Popović, 2015), ChrF++ (Popović, 2017), and TER (Snover et al., 2006) for translation into the five target languages across both evaluation sets in Table 15. By comparing the fourth and second

[12] https://huggingface.co/blog/constrained-beam-search

For the following translation into {tgt_lang}, please use the specified {tgt_lang} terms for the corresponding {src_lang} terms, while keeping the other content unchanged.

Term dictionary:
{terms}

{src_lang} text:
{src_text}

{tgt_lang} translation:
{tgt_text}

If multiple terms are nested or overlap with the context in {src_lang}, select the longest span that matches the context. Additionally, if a term has multiple meanings, only replace the term if its original context is relevant to the AI field. Provided the updated translation only.

Figure 12: Prompts for refining a machine translation model's initial output using relevant term translations from GIST as input.

You are tasked with improving the quality of translations containing AI-related terminologies. Given an English text and its machine-translated output in {tgt_lang}, your role is to analyze the translation for grammatical and morphological errors, such as incorrect word forms, cases, or agreement issues. You cannot change the content or meaning of the text; your task is strictly limited to fixing grammatical issues and ensuring proper linguistic structure.

Term dictionary:
{terms}

{src_lang} text:
{src_text}

{tgt_lang} translation:
{tgt_text}

Provided the updated translation only.

Figure 13: Prompts for refining the output of the word alignment method to ensure morphological coherence and accuracy.

values in each column of Table 2, we observe that the post-hoc prompting ablation enhances translation scores, aligning them with the performance of the prompting-powered method across all languages and models.

We also present one good example and one bad example for the prompting-powered refinement and the word alignment method across all languages, as shown from Figure 14 to Figure 18.

D.2 Qualitative Results

We also conduct a qualitative manual evaluation of the generation results produced by constrained beam search and logit adjustment methods, as explained in Appendix C.3. Both approaches exhibit extremely slow performance, running approximately 100 times slower than post-hoc methods. Furthermore, forcing specific word ids to appear in the output did not perform well, suffering from similar issues as the word alignment method in disrupting syntactical dependencies and morphological agreements. During manual inspection, we observe that constrained beam search behaves similarly to the logit adjustment method with a large scaling factor: the generated sentences often failed to maintain proper grammar, and terminological terms were frequently repeated multiple times within the same output. Conversely, with a small scaling factor in the logit adjustment method, the terms were often

omitted entirely, demonstrating little to no effect on the output.

D.3 Statistical Tests for Experiment Results

To assess the statistical significance of our findings in Section 6.3, we conducted one-sided paired t-tests on the BLEU, COMET, ChrF, ChrF++, and TER scores. Based on our observations, we formulated the following three hypotheses:

$H_0^{(1)}$: The prompting-powered refinement method outperforms direct translation across all languages.

$H_0^{(2)}$: The word alignment method outperforms direct translation for Chinese and Japanese but underperforms for Arabic, French, and Russian.

$H_0^{(3)}$: The prompting-powered refinement method outperforms the word alignment method across all languages.

As shown in Table 14, hypotheses 1 and 3 are fully supported by all metrics across all languages. Hypothesis 2 is partially supported, as there are cases where we cannot reject the null hypothesis that the word alignment method underperforms default translation for Arabic, French, and Russian. This outcome is insightful, as it suggests that the default translation approach remains effective for these languages.

Rank	Terminology	Rank1	Terminology (Continued)1	Rank2	Terminology (Continued)2
1	Algorithm	51	Latent space	101	Recall
2	Model	52	Node	102	Q-learning
3	Classifier	53	Transfer learning	103	Lasso
4	Transformer	54	Stochastic gradient descent	104	Transition matrix
5	Machine learning	55	Feature vector	105	Linear regression
6	Policy	56	Gibbs sampling	106	Meta-learning
7	Learning rate	57	Baseline	107	Segmentation
8	Neural network	58	Generative model	108	Fourier transform
9	Language model	59	Ontology	109	Epoch
10	Loss function	60	Attention	110	Learning algorithm
11	Reinforcement learning	61	Training set	111	Topic model
12	Encoder	62	Data mining	112	Time complexity
13	Deep learning	63	Manifold	113	Feature selection
14	Decoder	64	Discriminator	114	Knowledge distillation
15	Gradient descent	65	F1 score	115	Word embedding
16	Beam search	66	Dynamic programming	116	Euclidean distance
17	Machine translation	67	Adam optimizer	117	Covariance
18	Computer vision	68	Eigenvalue	118	Hyper-parameter
19	Dataset	69	Vector	119	Test set
20	Graph	70	State-of-the-art	120	Attention mechanism
21	Markov chain	71	Regularization	121	Oracle
22	Kernel	72	Backpropagation	122	Question answering
23	Marginal likelihood	73	Greedy algorithm	123	Point cloud
24	Objective function	74	Optical flow	124	Local minima
25	Gradient	75	Mutual information	125	N-gram
26	Reward function	76	Weight decay	126	Semi-supervised learning
27	Entropy	77	Posterior distribution	127	Batch normalization
28	Tensor	78	Bounding box	128	Homomorphism
29	Active learning	79	Disentanglement	129	Markov
30	Natural language processing	80	Convolution	130	Mini-batch
31	Perplexity	81	Semantic segmentation	131	Subgraph
32	Posterior	82	Logit	132	Bias
33	Logistic regression	83	Loss	133	Arg min
34	Covariance matrix	84	Multi-task learning	134	State space
35	Self-attention	85	Matrix	135	Dimensionality
36	Data augmentation	86	Binary classification	136	Random variable
37	Object detection	87	In-context learning	137	Gaussian distribution
38	Inference	88	Validation set	138	Optimizer
39	Cosine similarity	89	Cost function	139	Weight vector
40	Sample complexity	90	Corpus	140	Named entity recognition
41	Value function	91	Estimator	141	Kernel matrix
42	Probability distribution	92	Lemma	142	Discount factor
43	Generator	93	Parser	143	Hidden layer
44	Adam	94	Feature space	144	Domain adaptation
45	Supervised learning	95	Sentiment analysis	145	Frobenius norm
46	Dropout	96	Token	146	Positional encoding
47	Classification	97	Unsupervised learning	147	Seq2seq
48	Ground truth	98	State	148	Cross validation
49	Arg max	99	Infoset	149	Gaussian process
50	K-means	100	Precision	150	Coreference resolution

Table 5: Top 150 terms with the highest frequency in the original set of awarded papers.

Year	Terms
2000	feature vector, machine translation, probabilistic model, inference, loss function, feature space, BLEU, translation model, model selection, OOV
2001	reinforcement learning, policy, Ablation study, baseline, machine learning, bias, learning rate, state, gradient descent, tensor
2002	machine learning, machine translation, posterior distribution, inference, weight vector, training set, time complexity, tf-idf, mutual information, dot product
2003	softmax, computer vision, machine learning, local minima, gradient descent, neural network, Dataset, generative model, Gaussian noise, time complexity
2004	deep learning, ground truth, neural network, bounding box, 3D object detection, mean square error, machine learning, cost volume, gradient, receptive field
2005	neural network, loss function, machine translation, learning rate, gradient, perplexity, Kernel, validation set, beam search, feature vector
2006	NLP, machine learning, beam search, language model, Question Answering, Dataset, regularization, overfitting, greedy algorithm, neural network
2007	learning rate, probability distribution, machine learning, deep learning, loss, binary classification, covariance matrix, Adam optimizer, corpus, estimator
2008	machine learning, softmax, dropout, language model, computer vision, bias, baseline, neural network, F1 score, NLP
2009	computer vision, language model, softmax, NLP, deep learning, learning rate, clustering, natural language processing, neural network, loss function
2010	machine learning, learning rate, neural network, gradient descent, Euclidean distance, regularization, convex optimization, learning algorithm, Gaussian noise, gradient
2011	learning rate, NLP, deep learning, overfitting, supervised learning, Lemma, inference, eigenvalue, baseline, Cross Entropy Loss
2012	softmax, NLP, neural network, attention, deep learning, dropout, State-of-the-art, word embedding, learning rate, attention mechanism
2013	machine learning, loss function, Dataset, inference, NLP, learning rate, language model, stochastic gradient descent, probability distribution, decoder
2014	language model, NLP, machine learning, computer vision, convolutional layer, learning rate, deep learning, loss function, softmax, sigmoid
2015	NLP, language model, learning rate, feature vector, baseline, loss function, recall, validation set, probabilistic model, Markov chain
2016	learning rate, decoder, classification, loss function, NLP, Dataset, data augmentation, machine learning, State-of-the-art, beam search
2017	unsupervised learning, supervised learning, neural network, reinforcement learning, Dataset, dynamic programming, inference, validation set, NLP, epoch
2018	NLP, neural network, learning rate, machine learning, language model, Dataset, inference, natural language processing, loss function, BERT
2019	machine learning, learning rate, NLP, neural network, softmax, gradient descent, natural language processing, computer vision, node, Dataset
2020	NLP, Dataset, loss function, language model, natural language processing, learning rate, reinforcement learning, BERT, Adam optimizer, baseline
2021	learning rate, NLP, natural language processing, machine learning, language model, loss function, computer vision, BERT, overfitting, decoder
2022	machine learning, neural network, NLP, learning rate, computer vision, Dataset, loss function, deep learning, supervised learning, node
2023	machine learning, NLP, learning rate, computer vision, deep learning, Dataset, softmax, machine translation, natural language processing, neural network

Table 6: Top 10 most frequent terms in awarded papers for each year from 2000 to 2023.

Category	Selected Conferences
AI	AAAI, IJCAI
CV	CVPR, ECCV, ICCV
ML	ICLR, ICML, NeurIPS, KDD
NLP	ACL, EMNLP, NAACL, EACL LREC, COLING, CoNLL
Web & IR	SIGIR, WWW

Table 7: Top AI conferences included in our dataset collection.

	Arabic	Chinese	French	Japanese	Russian
Three-Model Agreement Ratio	10.11%	42.71%	9.86%	16.60%	17.93%
Two-Model Agreement Ratio	30.52%	36.82%	45.15%	40.44%	38.36%

Table 8: Translation agreement ratios among three models (Claude 3 Sonnet, Google Translate API, and GPT-3.5-Turbo) for five target languages. The table shows the ratio of terms where all three models agree (*Three-Model Agreement Ratio*) and the ratio where any two models agree (*Two-Model Agreement Ratio*).

Language	Claude 3 vs. Human	Google Translate vs. Human	GPT-3.5 vs. Human	GPT-4o vs. Human (500 Random Examples)
Chinese	69.26%	62.84%	59.20%	76.80%
Arabic	29.26%	36.45%	23.05%	39.80%
French	57.06%	51.44%	14.17%	58.20%
Japanese	57.59%	49.65%	34.07%	67.80%
Russian	39.72%	42.63%	28.23%	41.20%

Table 9: Comparison of translation accuracy between Claude 3, Google Translate, GPT-3.5, and GPT-4o against human translations across different languages.

	A. Both translations are good	B. Method 1 translation is better	C. Method 2 translation is better	D. Both translations are bad
Arabic	739, 45.76%	461, 28.54%	329, 20.37%	72, 4.46%
Chinese	468, 50.59%	266, 28.76%	163, 17.62%	25, 2.70%
French	438, 48.67%	274, 30.44%	170, 18.89%	17, 1.89%
Japanese	587, 56.99%	251, 24.37%	159, 15.44%	27, 2.62%
Russian	626, 54.43%	348, 30.26%	150, 13.04%	14, 1.22%

Table 10: Distribution of annotators' choices for AI terminology ratings in Task 1, comparing the GPT-4o-selected candidate (Method 1) with the majority-voted candidate (Method 2). The table reports the total counts and corresponding ratios for each choice.

	A. Both translations are good	B. Method 1 translation is better	C. Method 2 translation is better	D. Both translations are bad
Arabic	376, 46.42%	238, 29.38%	143, 17.65%	46, 5.68%
Chinese	197, 37.17%	228, 43.02%	85, 16.04%	17, 3.21%
French	152, 39.48%	168, 43.64%	53, 13.77%	10, 2.60%
Japanese	295, 57.28%	162, 31.46%	36, 6.99%	21, 4.08%
Russian	172, 39.09%	198, 45.00%	38, 8.64%	25, 5.68%

Table 11: Distribution of annotators' choices for AI terminology ratings in Task 2, comparing the translations in our dataset (Method 1) with those in the 60-60 initiative evaluation set (Method 2). The table reports the total counts and corresponding ratios for each choice.

	Arabic	Chinese	French	Japanese	Russian
Task 1	0.30	0.22	0.37	0.21	0.39
Task 2	0.22	0.41	0.50	0.39	0.41

Table 12: Fleiss' Kappa scores for inter-annotator agreement among 5 annotators across each question in Task 1 and Task 2.

Language	**Task 1 (Mean $\pm$ Std)**	**Task 2 (Accuracy)**	**Task 3 (Accuracy)**
Chinese	3.96 ± 1.03	0.88 ± 0.32	0.91 ± 0.29
Arabic	3.39 ± 1.27	0.90 ± 0.30	0.91 ± 0.28
French	3.81 ± 1.20	0.78 ± 0.42	0.80 ± 0.40
Japanese	3.24 ± 1.13	0.77 ± 0.42	0.72 ± 0.45
Russian	3.11 ± 1.55	0.83 ± 0.38	0.82 ± 0.39

Table 13: Manual evaluation results for translation quality, grammatical correctness, and terminology accuracy across five languages.

		Arabic		Chinese		French		Japanese		Russian	
Metric	Hypothesis	T Stats	P Val	T Stats	P Val	T Stats	P Val	T Stats	P Val	T Stats	P Val
COMET	1	14.18	0.00	15.22	0.00	12.68	0.00	11.13	0.00	8.53	0.00
	2	6.91	0.00	7.60	0.00	1.84	0.03	3.21	0.00	20.76	0.00
	3	17.93	0.00	11.67	0.00	13.58	0.00	10.62	0.00	24.89	0.00
BLEU	1	6.37	0.00	6.72	0.00	9.80	0.00	6.26	0.00	7.63	0.00
	2	-0.66	0.75	3.89	0.00	-0.22	0.59	2.58	0.00	-2.85	1.00
	3	4.80	0.00	3.87	0.00	8.77	0.00	4.57	0.00	5.50	0.00
ChrF	1	6.37	0.00	6.72	0.00	9.79	0.00	6.26	0.00	7.64	0.00
	2	-0.66	0.75	3.89	0.00	-0.18	0.57	2.59	0.00	-2.85	1.00
	3	4.79	0.00	3.86	0.00	8.76	0.00	4.55	0.00	5.49	0.00
ChrF++	1	6.37	0.00	6.72	0.00	9.79	0.00	6.26	0.00	7.64	0.00
	2	-0.66	0.75	3.89	0.00	-0.18	0.57	2.59	0.00	-2.85	1.00
	3	4.79	0.00	3.86	0.00	8.76	0.00	4.55	0.00	5.49	0.00
TER	1	6.17	0.00	6.72	0.00	9.71	0.00	6.26	0.00	7.41	0.00
	2	-0.85	0.80	3.88	0.00	-0.18	0.57	2.59	0.00	-2.76	1.00
	3	4.40	0.00	3.87	0.00	8.73	0.00	4.56	0.00	5.27	0.00

Table 14: Hypothesis test statistics and p-values for all metrics across the five tested models.

Model	Metric	Arabic	Chinese	French	Japanese	Russian
Evaluation Set: 60-60						
aya-expanse	BLEU	20.11 + 1.23 + 0.18 + 1.24	27.31 + 1.33 + 0.24 + 0.53	33.05 + 2.46 + 0.20 + 1.97	14.59 + 0.61 + 0.32 + 1.08	16.59 + 1.59 - 0.05 + 0.92
	ChrF	20.62 + 1.24 + 0.18 + 1.26	27.52 + 1.32 + 0.24 + 0.52	33.68 + 2.44 + 0.18 + 1.96	14.76 + 0.61 + 0.32 + 1.08	16.99 + 1.59 - 0.06 + 0.92
	ChrF++	20.62 + 1.24 + 0.18 + 1.26	27.52 + 1.32 + 0.24 + 0.52	33.68 + 2.44 + 0.18 + 1.96	14.76 + 0.61 + 0.32 + 1.08	16.99 + 1.59 - 0.06 + 0.92
	COMET	81.96 + 0.71 - 0.52 + 1.27	83.43 + 1.57 + 0.08 + 1.63	81.83 + 1.06 - 0.11 + 1.35	88.54 + 0.32 - 0.01 + 0.38	82.27 + 0.69 - 2.02 + 0.98
	TER	93.61 - 1.35 - 0.24 - 1.36	73.47 - 1.33 - 0.24 - 0.53	77.43 - 2.78 - 0.23 - 2.28	86.42 - 0.61 - 0.33 - 1.08	94.56 - 1.74 + 0.08 - 0.96
aya-23-8B	BLEU	19.98 + 0.54 - 0.21 + 0.58	26.08 + 0.47 + 0.39 + 0.45	33.85 + 2.28 - 0.11 + 2.48	15.06 + 0.87 + 0.36 + 1.24	15.77 + 1.05 + 0.37 + 0.84
	ChrF	20.50 + 0.54 - 0.21 + 0.58	26.20 + 0.47 + 0.39 + 0.45	34.48 + 2.26 - 0.11 + 2.46	15.22 + 0.87 + 0.37 + 1.24	16.17 + 1.06 + 0.37 + 0.85
	ChrF++	20.50 + 0.54 - 0.21 + 0.58	26.20 + 0.47 + 0.39 + 0.45	34.48 + 2.26 - 0.11 + 2.46	15.22 + 0.87 + 0.37 + 1.24	16.17 + 1.06 + 0.37 + 0.85
	COMET	84.02 + 0.81 - 0.24 + 1.05	85.12 + 0.58 + 0.38 + 0.93	82.40 + 0.94 - 0.15 + 1.26	87.92 + 0.50 + 0.09 + 0.70	81.91 + 0.40 - 2.26 + 0.84
	TER	93.97 - 0.57 + 0.23 - 0.61	74.40 - 0.46 - 0.38 - 0.44	76.35 - 2.64 + 0.18 - 2.89	85.73 - 0.88 - 0.37 - 1.24	95.60 - 1.17 - 0.45 - 0.98
gpt-4o-mini	BLEU	23.58 + 1.07 - 0.00 + 0.50	32.64 + 1.60 + 0.66 + 1.48	40.80 + 3.08 + 0.50 + 2.38	21.46 + 0.64 + 0.19 + 0.94	17.25 + 1.07 - 0.13 + 0.56
	ChrF	24.06 + 1.07 - 0.00 + 0.50	32.76 + 1.60 + 0.66 + 1.48	41.40 + 3.06 + 0.49 + 2.37	21.59 + 0.64 + 0.19 + 0.94	17.65 + 1.08 - 0.13 + 0.56
	ChrF++	24.06 + 1.07 - 0.00 + 0.50	32.76 + 1.60 + 0.66 + 1.48	41.40 + 3.06 + 0.49 + 2.37	21.59 + 0.64 + 0.19 + 0.94	17.65 + 1.08 - 0.13 + 0.56
	COMET	85.77 + 0.69 - 0.44 + 0.61	87.30 + 0.48 + 0.26 + 0.44	84.56 + 0.68 - 0.04 + 0.62	89.96 + 0.14 + 0.01 + 0.14	83.68 + 0.38 - 2.29 + 0.47
	TER	89.66 - 1.17 - 0.11 - 0.54	67.79 - 1.60 - 0.66 - 1.48	68.32 - 3.57 - 0.56 - 2.76	79.33 - 0.64 - 0.20 - 0.94	93.90 - 1.17 + 0.15 - 0.58
nllb	BLEU	22.38 + 1.37 + 0.64 + 1.21	17.29 + 1.92 + 1.02 + 2.40	34.93 + 2.86 + 0.21 + 3.23	6.19 + 2.42 + 0.53 + 2.95	17.30 + 1.54 + 1.07 + 1.51
	ChrF	22.87 + 1.37 + 0.64 + 1.22	17.45 + 1.92 + 1.03 + 2.40	35.55 + 2.85 + 0.21 + 3.21	6.22 + 2.42 + 0.53 + 2.94	17.69 + 1.55 + 1.08 + 1.53
	ChrF++	22.87 + 1.37 + 0.64 + 1.22	17.45 + 1.92 + 1.03 + 2.40	35.55 + 2.85 + 0.21 + 3.21	6.22 + 2.42 + 0.53 + 2.94	17.69 + 1.55 + 1.08 + 1.53
	COMET	83.52 + 0.83 - 0.45 + 1.31	78.22 + 2.95 + 0.73 + 3.67	82.83 + 1.00 - 0.19 + 1.43	77.82 + 3.80 + 0.39 + 4.80	81.41 + 0.97 - 1.54 + 1.79
	TER	91.12 - 1.59 - 0.74 - 1.34	83.26 - 1.93 - 1.03 - 2.40	75.21 - 3.24 - 0.26 - 3.72	94.32 - 2.41 - 0.53 - 2.94	93.87 - 1.67 - 1.17 - 1.65
seamless	BLEU	23.13 + 1.16 - 0.03 + 2.35	26.26 + 0.97 + 0.80 + 2.75	40.04 + 2.08 - 0.57 + 1.69	14.56 + 0.74 + 0.05 + 2.38	17.18 + 1.71 + 1.17 + 1.22
	ChrF	23.67 + 1.17 - 0.03 + 2.36	26.43 + 0.98 + 0.80 + 2.74	40.68 + 2.08 - 0.58 + 1.67	14.65 + 0.74 + 0.05 + 2.37	17.61 + 1.71 + 1.17 + 1.23
	ChrF++	23.67 + 1.17 - 0.03 + 2.36	26.43 + 0.98 + 0.80 + 2.74	40.68 + 2.08 - 0.58 + 1.67	14.65 + 0.74 + 0.05 + 2.37	17.61 + 1.71 + 1.17 + 1.23
	COMET	84.07 + 0.94 - 0.38 + 1.26	83.44 + 1.48 + 0.50 + 2.49	83.86 + 0.78 - 0.07 + 1.05	85.05 + 1.06 + 0.16 + 1.74	82.33 + 0.56 - 1.87 + 1.21
	TER	90.27 - 1.33 + 0.01 - 2.78	74.32 - 0.97 - 0.80 - 2.74	69.19 - 2.38 + 0.65 - 1.91	86.02 - 0.74 - 0.05 - 2.38	94.01 - 1.89 - 1.30 - 1.38
Evaluation Set: AI Papers & Model Cards						
aya-expanse	BLEU	11.47 + 0.37 + 0.10 + 0.49	12.04 + 0.94 + 0.17 + 0.85	18.84 + 1.71 - 0.85 + 2.19	8.11 - 0.03 + 0.04 + 0.31	13.84 + 0.21 + 0.32 + 0.68
	ChrF	11.95 + 0.37 + 0.10 + 0.50	12.76 + 0.98 + 0.18 + 0.87	19.49 + 1.74 - 0.85 + 2.22	8.59 - 0.02 + 0.06 + 0.33	14.32 + 0.21 + 0.34 + 0.70
	ChrF++	11.95 + 0.37 + 0.10 + 0.50	12.76 + 0.98 + 0.18 + 0.87	19.49 + 1.74 - 0.85 + 2.22	8.59 - 0.02 + 0.06 + 0.33	14.32 + 0.21 + 0.34 + 0.70
	COMET	80.98 + 0.46 - 0.51 + 0.63	82.42 + 0.56 - 0.01 + 0.75	81.16 + 0.36 - 0.79 + 0.44	85.48 + 0.34 + 0.07 + 0.54	82.76 + 0.47 - 2.10 + 0.75
	ter	106.35 - 0.30 - 0.13 - 0.47	99.34 - 1.04 - 0.21 - 0.94	96.12 - 2.02 + 0.99 - 2.49	101.99 - 0.03 - 0.07 - 0.39	99.77 - 0.19 - 0.40 - 0.79
aya-23-8B	BLEU	14.28 + 0.81 + 0.28 + 0.75	14.50 + 0.39 + 0.19 + 0.71	24.49 + 2.36 + 0.08 + 2.37	9.22 + 0.23 + 0.35 + 0.45	16.39 + 1.36 + 0.61 + 1.35
	ChrF	14.77 + 0.82 + 0.28 + 0.75	15.12 + 0.40 + 0.20 + 0.71	25.11 + 2.37 + 0.07 + 2.39	9.68 + 0.25 + 0.37 + 0.47	16.78 + 1.39 + 0.62 + 1.37
	ChrF++	14.77 + 0.82 + 0.28 + 0.75	15.12 + 0.40 + 0.20 + 0.71	25.11 + 2.37 + 0.07 + 2.39	9.68 + 0.25 + 0.37 + 0.47	16.78 + 1.39 + 0.62 + 1.37
	COMET	81.55 + 1.03 - 0.62 + 1.23	83.88 + 0.68 + 0.04 + 0.80	82.55 + 1.22 - 0.70 + 1.23	84.42 + 0.81 + 0.02 + 1.26	82.72 + 1.14 - 2.08 + 1.29
	ter	102.45 - 0.92 - 0.37 - 0.91	95.18 - 0.69 - 0.42 - 0.97	89.07 - 2.93 - 0.28 - 2.99	100.30 - 0.43 - 0.58 - 0.72	96.28 - 1.73 - 0.84 - 1.69
gpt-4o-mini	BLEU	14.37 + 0.53 - 0.42 + 0.61	17.21 + 1.22 + 1.39 + 1.34	24.45 + 4.38 + 1.28 + 3.40	10.55 + 0.05 + 0.03 + 0.01	18.02 + 1.52 + 0.40 + 1.60
	ChrF	14.82 + 0.55 - 0.42 + 0.63	17.93 + 1.23 + 1.39 + 1.37	25.04 + 4.40 + 1.31 + 3.45	10.99 + 0.06 + 0.03 + 0.03	18.44 + 1.52 + 0.40 + 1.62
	ChrF++	14.82 + 0.55 - 0.42 + 0.63	17.93 + 1.23 + 1.39 + 1.37	25.04 + 4.40 + 1.31 + 3.45	10.99 + 0.06 + 0.03 + 0.03	18.44 + 1.52 + 0.40 + 1.62
	COMET	83.56 + 0.86 - 0.19 + 0.93	86.08 + 0.25 - 0.13 + 0.07	84.75 + 0.33 - 1.03 + 0.13	87.91 + 0.16 - 0.01 + 0.19	84.92 + 0.44 - 2.02 + 0.49
	ter	102.33 - 1.04 + 0.25 - 1.16	93.01 - 2.23 - 2.45 - 2.24	89.16 - 5.36 - 1.80 - 4.21	98.39 - 0.07 - 0.13 - 0.06	94.40 - 1.81 - 0.51 - 1.93
nllb	BLEU	15.42 - 0.31 - 0.77 + 0.29	10.24 + 2.19 + 2.07 + 3.10	22.68 + 2.73 + 0.90 + 3.40	8.24 + 1.08 + 0.88 + 1.21	19.18 - 0.10 - 0.24 + 1.44
	ChrF	15.95 - 0.32 - 0.77 + 0.30	10.82 + 2.26 + 2.16 + 3.21	23.33 + 2.77 + 0.91 + 3.43	8.61 + 1.11 + 0.93 + 1.27	19.63 - 0.07 - 0.20 + 1.50
	ChrF++	15.95 - 0.32 - 0.77 + 0.30	10.82 + 2.26 + 2.16 + 3.21	23.33 + 2.77 + 0.91 + 3.43	8.61 + 1.11 + 0.93 + 1.27	19.63 - 0.07 - 0.20 + 1.50
	COMET	81.23 + 1.28 - 0.50 + 1.99	80.19 + 1.61 + 0.38 + 2.84	78.70 + 3.81 - 1.59 + 1.62	83.05 + 1.71 + 0.70 + 2.55	80.46 + 2.80 - 2.38 + 1.63
	ter	101.29 + 0.17 + 0.71 - 0.47	101.05 - 2.45 - 2.59 - 3.46	91.40 - 3.07 - 0.89 - 4.01	101.41 - 1.31 - 0.86 - 1.28	93.26 + 0.27 + 0.37 - 1.58
seamless	BLEU	15.38 + 1.09 + 0.45 + 1.66	13.67 + 1.10 + 0.73 + 1.99	24.34 + 5.21 + 1.49 + 5.35	9.42 + 0.56 + 0.42 + 0.91	18.43 + 0.95 + 0.35 + 1.70
	ChrF	15.90 + 1.10 + 0.46 + 1.68	14.34 + 1.08 + 0.72 + 2.05	25.05 + 5.25 + 1.50 + 5.37	9.91 + 0.55 + 0.41 + 0.91	18.88 + 0.97 + 0.37 + 1.73
	ChrF++	15.90 + 1.10 + 0.46 + 1.68	14.34 + 1.08 + 0.72 + 2.05	25.05 + 5.25 + 1.50 + 5.37	9.91 + 0.55 + 0.41 + 0.91	18.88 + 0.97 + 0.37 + 1.73
	COMET	81.96 + 1.18 - 0.39 + 1.72	80.70 + 2.18 + 0.16 + 2.95	83.76 + 0.97 - 0.94 + 0.88	83.70 + 0.88 + 0.10 + 1.81	83.12 + 1.50 - 1.79 + 1.70
	ter	101.34 - 1.48 - 0.77 - 2.19	97.26 - 1.76 - 1.07 - 2.62	89.75 - 6.47 - 2.12 - 6.61	100.59 - 1.35 - 1.15 - 1.77	94.15 - 1.03 - 0.38 - 2.03

Table 15: Full evaluation results across five models and five languages using BLEU, ChrF, ChrF++, COMET, and TER. The first black value in each column represents direct translation scores. The second, third, and fourth values (colored red or green) indicate the relative change in performance when applying the prompting-powered refinement method, the word alignment method, and post-hoc prompting after word alignment for improved morphological and grammatical correctness, respectively, compared to direct translation. Lower TER scores indicate better alignment with the reference.

Method: Word Alignment; **Example**: Good, **COMET Score**: 0.9864
English Text: The provided references lack this granular information.
Ground Truth: المراجع المقدمة تفتقر إلى هذه المعلومات الدقيقة.
Method Translation: المراجع المقدمة تفتقر إلى هذه المعلومات الدقيقة.

Method: Word Alignment; **Example**: Bad, **COMET Score**: 0.3019
English Text: This work is done by the Data Science Applied Research team at LinkedIn. Special thanks to our TPM Shruti Sharma and our collaborators in Data Science, Engineering, SRE, FP\&A, BizOps, and Product for adopting the library. In particular , thanks to Ashok Sridhar , Mingyuan Zhong , Peter Huang , Hamin Oh , Neha Gupta , Neelima Rathi , Deepti Rai , Christian Rhally , Camilo Rivera , Priscilla Tam , Meenakshi Adaikalavan , Zheng Shao , Mike Snow , Stephen Bisordi , Dong Wang , Ankit Upadhyaya , and Rachit Kumar for allowing us to share their use cases
Ground Truth: شكر خاص لـ LinkedIn. تم تنفيذ هذا العمل بواسطة فريق الأبحاث التطبيقية لعلوم البيانات في TPM Shruti Sharma والمنتج لتبني المكتبة. على وجه الخصوص، شكرًا لأشوك سريدهار، BizOps، وFP\&A، وSRE والمتعاونين معنا في علوم البيانات، والهندسة، و مينجيوان تشونغ، بيتر هوانغ، هامين أوه، نيها غوبتا، نيليما راثي، ديبتي راي، كريستيان رهالي، كاميلو ريفيرا، بريسيلا تام، ميناكشي أديكالافان، تشنغ شاو، مايك سنو، ستيفن بيسوردي، دونغ وانغ، أنكيت أوبادهيايا، ورشيت كومار للسماح لنا لتبادل حالات الاستخدام الخاصة بهم
Method Translation: شكرًا خاصًا لشركة LinkedIn. يتم هذا العمل من قبل فريق البحوث التطبيقية في علوم البيانات في TPM Shruti Sharma ومتعاونينا في علوم البيانات والهندسة والهند

Method: Prompting Refinement; **Example**: Good, **COMET Score**: 0.9864
English Text: The provided references lack this granular information.
Ground Truth: المراجع المقدمة تفتقر إلى هذه المعلومات الدقيقة.
Method Translation: المراجع المقدمة تفتقر إلى هذه المعلومات الدقيقة.

Method: Prompting Refinement; **Example**: Bad, **COMET Score**: 0.5784
English Text: "We pre-train LiLT on the IIT-CDIP Test Collection 1.0 (Lewis et al., 2006), which is a large-scale scanned document image dataset and contains more than 6 million documents with more than 11 million scanned document images." So the software type that the model is trained on is scanned document images.
Ground Truth: "وهي عبارة عن مجموعة IIT-CDIP Test Collection 1.0 (Lewis et al., 2006)، لقد قمنا بتدريب LiLT مسبقًا على بيانات صور مستندات ممسوحة ضوئيًا واسعة النطاق وتحتوي على أكثر من 6 ملايين مستند مع أكثر من 11 مليون صورة مستندات ممسوحة ضوئيًا." لذا فإن نوع البرنامج الذي يتم تدريب النموذج عليه هو صور المستندات الممسوحة ضوئيًا.
Method Translation: "اختبار المجموعة 1.0 (ليوس وآخرون ، 2006) ، وهو على نطاق واسع IIT-CDIP نحن التدريب المسبق LiLT على مسح صورة المستند مجموعة بيانات ويحتوي على أكثر من 6 ملايين وثيقة مع أكثر من 11 مليون مسح صورة الوثيقة".

Figure 14: Arabic translation examples of the two integration cases on `nllb`, including one good example and one bad example for each case.

Method: Word Alignment; **Example**: Good, **COMET Score**: 0.9676
English Text: Table 8 provides English pronouns and the list of possible target pronouns.
Ground Truth: 表 8 提供了英语代词和可能的目标代词列表。
Method Translation: 表8提供了英语代词和可能的目标代词列表.

Method: Word Alignment; **Example**: Bad, **COMET Score**: 0.3132
English Text: They provide inmates the ability to make and receive unauthorized phone calls, send email and text messages, use social media, and follow news pertaining to their case, among other forbidden uses.
Ground Truth: 它们使囚犯能够拨打和接听未经授权的电话、发送电子邮件和短信、使用社交媒体、关注与其案件有关的新闻以及其他禁止的用途。
Method Translation: 监狱的监狱和监狱的监狱都在使用这些网络,

Method: Prompting Refinement; **Example**: Good, **COMET Score**: 0.9609
English Text: The provided references lack this granular information.
Ground Truth: 提供的参考文献缺乏这种详细信息。
Method Translation: 提供的参考文献没有这些细节信息.

Method: Prompting Refinement; **Example**: Bad, **COMET Score**: 0.3132
English Text: They provide inmates the ability to make and receive unauthorized phone calls, send email and text messages, use social media, and follow news pertaining to their case, among other forbidden uses.
Ground Truth: 它们使囚犯能够拨打和接听未经授权的电话、发送电子邮件和短信、使用社交媒体、关注与其案件有关的新闻以及其他禁止的用途。
Method Translation: 监狱的监狱和监狱的监狱都在使用这些网络,

Figure 15: Chinese translation examples of the two integration cases on `nllb`, including one good example and one bad example for each case.

Method: Word Alignment; **Example**: Good, **COMET Score**: 0.9749
English Text: Nevertheless, we gratefully acknowledge that our work was inspired by Shannon's original idea.
Ground Truth: Néanmoins, nous reconnaissons avec gratitude que notre travail a été inspiré par l'idée originale de Shannon.
Method Translation: Néanmoins, nous reconnaissons avec gratitude que notre travail a été inspiré par l'idée originale de Shannon.

Method: Word Alignment; **Example**: Bad, **COMET Score**: 0.2879
English Text: Let a i denote the total noise in the accumulator at the end of iteration i. Let c i and u i denote the amount of correlated and uncorrelated noise, respectively, on iteration i and let r i denote the rank.
Ground Truth: Soit a i le bruit total dans l'accumulateur à la fin de l'itération i. Soient c i et u i la quantité de bruit corrélé et non corrélé, respectivement, lors de l'itération i et r i désigne le rang.
Method Translation: La quantité de bruit correlate est calculée en fonction de la itération de la fréquence de

Method: Prompting Refinement; **Example**: Good, **COMET Score**: 0.9602
English Text: [More Information Needed] on how this comparison could be leveraged for explainability or interpretability purposes.
Ground Truth: [Plus d'informations nécessaires] sur la manière dont cette comparaison pourrait être exploitée à des fins d'explicabilité ou d'interprétabilité.
Method Translation: [Plus d'informations nécessaires] sur la manière dont cette comparaison pourrait être exploitée à des fins d'explicabilité ou d'interprétabilité.

Method: Prompting Refinement; **Example**: Bad, **COMET Score**: 0.3012
English Text: r π * − g + σ P π * (V) − V = max π {r π − g + σ P π (V) − V } . (129)

Ground Truth: r π * − g + σ P π * (V) − V = max π {r π − g + σ P π (V) − V } . (129)
Method Translation: Le nombre de points de référence est le nombre de points de référence de la valeur de base.

Figure 16: French translation examples of the two integration cases on `nllb`, including one good example and one bad example for each case.

Method: Word Alignment; **Example**: Good, **COMET Score**: 0.9542
English Text: This section provides an analysis of our results at each step of our experiment. First, we investigate the results on the dev dataset which guided our model selection for the evaluation phase.
Ground Truth: このセクションでは、実験の各ステップでの結果の分析を提供します。まず、評価フェーズのモデル選択のガイドとなった開発データセットの結果を調査します。
Method Translation: このセクションでは,実験の各段階における結果の分析を提供します.まず,評価段階のモデル選択を導く開発データセットの結果を調査します.

Method: Word Alignment; **Example**: Bad, **COMET Score**: 0.2297
English Text: [2003] optimize a CPG-based (central pattern generator) locomotion controller [Taga 1995] for 3D musculoskeletal models without tendon or activation dynamics, but their results were not compared to human kinematic and dynamic gait patterns. Moreover, full musculoskeletal models are significantly more difficult to construct than joint-actuated models.
Ground Truth: [2003] は、腱や活性化ダイナミクスを使用しない 3D 筋骨格モデル用に CPG ベース（中央パターン ジェネレーター）の移動コントローラー [Taga 1995] を最適化しましたが、その結果は人間の運動学的および動的歩行パターンと比較されていませんでした。さらに、完全な筋骨格モデルは、関節作動モデルよりも構築が大幅に困難です。
Method Translation: [2003] 動力制御装置 (CPG) ベースの運動制御装置 (中央パターン生成ジェネレーター) [Taga 1995] を動脈

Method: Prompting Refinement; **Example**: Good, **COMET Score**: 0.9608
English Text: Here, we show that Theorem 1 can be generalized to products of arbitrary permutation groups.
Ground Truth: ここで、定理 1 が任意の置換群の積に一般化できることを示します。
Method Translation: ここでは定理1が任意の順列群の積に一般化できることを示しています。

Method: Prompting Refinement; **Example**: Bad, **COMET Score**: 0.5776
English Text: . , z D ~ N (0, 1), P r D d=1 u d z d ≥ u 2 • c 4 ln 2 δ ≤ δ.
E Proof for Theorem 3.1
In this section, we prove the failure of uniform convergence for our linear model. We first recall the setup:
Ground Truth: 。 , z D ∽ N (0, 1), P r D d=1 u d z d ≥ u 2 ・ c 4 ln 2 δ ≤ δ。
E 定理 3.1 の証明
このセクションでは、線形モデルの一様収束が失敗することを証明します。まずセットアップを思い出します。
Method Translation: 定理の証明 3.1 このセクションでは,我々の線形モデルにおける一様収束の失敗を証明する.まず,設定を再現率します.

Figure 17: Japanese translation examples of the two integration cases on `nllb`, including one good example and one bad example for each case.

Method: Word Alignment; **Example**: Good, **COMET Score**: 0.9735
English Text: Using this approach, a BRDF can be measured in just a few minutes.
Ground Truth: Используя этот подход, BRDF можно измерить всего за несколько минут.
Method Translation: Используя этот подход, BRDF можно измерить всего за несколько минут.

Method: Word Alignment; **Example**: Bad, **COMET Score**: 0.5001
English Text: The AUDIT algorithm for auditing the full recommender system is described in Alg. 2. AUDIT samples a subset of users and a subset of arms for each sampled user. Then it applies OCEF to each user simultaneously Algorithm 2: AUDIT algorithm.
Ground Truth: Алгоритм AUDIT для аудита всей рекомендательной системы описан в Алг. 2. AUDIT осуществляет выборку подмножества пользователей и подмножества рук для каждого отобранного пользователя. Затем он применяет OCEF к каждому пользователю одновременно. Алгоритм 2: Алгоритм AUDIT.
Method Translation: Алгоритм AUDIT для аудита полной рекомендательная система описан в ал. 2.

Method: Prompting Refinement; **Example**: Good, **COMET Score**: 0.9712
English Text: The experiment was implemented in Python 3.6 and the packages listed in table 11
Ground Truth: Эксперимент был реализован на Python 3.6 и пакетах, перечисленных в таблице 11.
Method Translation: Эксперимент был реализован в Python 3.6 и пакетах, перечисленных в таблице 11.

Method: Prompting Refinement; **Example**: Bad, **COMET Score**: 0.5820
English Text: Guy hurriedly leaves the compartment but leaves Bruno thinking he has agreed to the deal . Guy accidentally leaves his cigarette lighter behind , a gift from Anne to Guy , Which Bruno pockets .
Ground Truth: Гай поспешно покидает купе, но Бруно думает, что тот согласился на сделку. Гай случайно оставляет свою зажигалку, подарок Анны Гаю, который Бруно кладет в карман.
Method Translation: Гай поспешно покидает купе, но бросает Бруно, думая, что он согласился на сделку.

Figure 18: Russian translation examples of the two integration cases on `nllb`, including one good example and one bad example for each case.