ACADATA: Parallel Dataset of Academic Data for Machine Translation

Iñaki Lacunza[*1], Javier Garcia Gilabert[*1], Francesca De Luca Fornaciari[*1], Javier Aula-Blasco[1], Aitor Gonzalez-Agirre[1], Maite Melero[1], Marta Villegas[1]
[1]Barcelona Supercomputing Center (BSC)
inaki.lacunza@bsc.es, javier.garcia1@bsc.es, fdelucaf@bsc.es

Abstract

We present ACADATA, a high-quality parallel dataset for academic translation, that consists of two subsets: ACAD-TRAIN, which contains approximately 1.5 million human-generated paragraph pairs across 12 languages, and ACAD-BENCH, a curated evaluation set of almost 6,000 translations covering 12 directions. To validate its usefulness, we fine-tune two Large Language Models (LLMs) on ACAD-TRAIN and benchmark them on ACAD-BENCH against specialized machine-translation systems, general-purpose, open-weight LLMs, and several large-scale proprietary models. Experimental results demonstrate that fine-tuning on ACAD-TRAIN leads to improvements in academic translation quality by +6.1 and +12.4 d-BLEU points on average for 7B and 2B models respectively, while also improving long-context translation in a general domain by up to 24.9% when translating out of English. The fine-tuned top-performing model surpasses the best proprietary and open-weight models on the academic translation domain. By releasing ACAD-TRAIN, ACAD-BENCH and the fine-tuned models, we provide the community with a valuable resource to advance research in the academic domain and long-context translation.

Keywords: Academic Translation, Multilingual Dataset, Machine Translation, Parallel Corpus

1. Introduction

While English has been long established as the lingua franca for scientific research, a significant volume of impactful work is being published in other languages (Stockemer and Wigginton, 2019). As a result, Machine Translation (MT) has become essential for extending access to and integrating academic findings. However, building MT systems tailored to the academic domain presents challenges that differ from general-purpose translation (e.g., domain-specific terminology, emerging neologisms, complex syntactic constructions), often leading to reduced translation quality (Roussis et al., 2024).

Training state-of-the-art MT systems for less-represented languages and specialized domains requires large amounts of high-quality parallel data. In fact, when training translation systems, the quality of parallel corpora plays a critical role not only during the different stages of model training (e.g. pre-training, supervised fine-tuning, preference optimization, etc.) but also in evaluation. Yet, constructing such corpora is far from easy.

Inspired by prior work leveraging parallel text from public academic repositories, we introduce the ACADATA dataset: a multilingual parallel corpus extracted from academic abstracts made openly accessible by various research institutions. The dataset contains paragraph-level translations for a wide set of European language pairs in the academic domain, covering specialized content across a broad range of disciplines. All texts are collected from the metadata accompanying the published

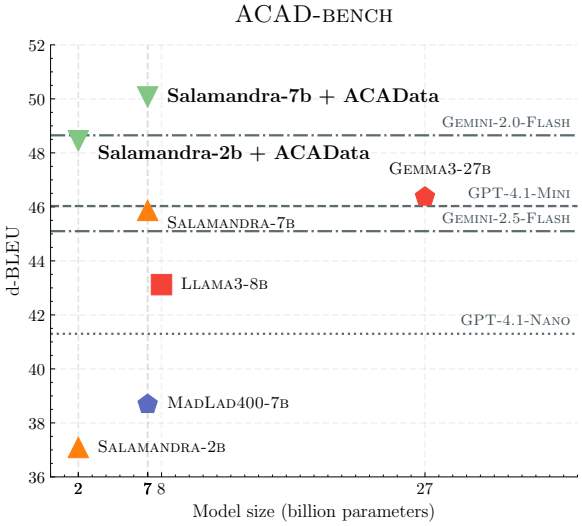

Figure 1: Translation quality on ACAD-BENCH in xx→en directions for models fine-tuned with ACAD-TRAIN and a set of open-weight and proprietary systems of different scales. When the scale is not known, we represent it with a horizontal line.

works and consist of firsthand translations provided by their authors. ACADATA is divided into two subsets[1]

- **ACAD-TRAIN:** a high-quality parallel training dataset of 1,461,418 paragraph translations across 12 European languages. This human-generated, academic domain-specific data

* Core contributors.

[1]https://huggingface.co/datasets/BSC-LT/ACAData

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8498–8519
11–16 May 2026.

has been obtained from multiple research institutions.

- **ACAD-BENCH:** an evaluation set of 5,944 translation instances for benchmarking MT systems on formal, technical, and scientific domains. This test set covers the most frequent language pairs present in the training corpus.

We demonstrate the usefulness of the ACAD-TRAIN dataset by fine-tuning two LLMs and evaluating their performance on ACAD-BENCH before and after fine-tuning[2][3], as well as in the standard general-domain WMT24++ dataset. Additionally, we benchmark a range of open-weight and proprietary systems on ACAD-BENCH to assess their academic specific translation capabilities and provide some reference results on our benchmarking split. Despite their relatively small size, the fine-tuned models achieve a level of translation quality that is on a par with, or even surpasses, that of proprietary models on ACAD-BENCH (see Figure 1).

The entire ACADATA dataset statistics are summarized in Table 1. By open-sourcing our dataset (CC BY 4.0 license) and fine-tuned models (Apache 2.0 license), we encourage the machine translation community to pursue research on academic domain translation.

	ACAD-TRAIN	ACAD-BENCH
Instances	1,461,418	5,944
Languages	12	5
Directions	96	12
Src len ($\mu \pm \sigma$)	1,051 ± 759	1,091 ± 779
Tgt len ($\mu \pm \sigma$)	1,107 ± 814	1,169 ± 832

Table 1: Summary statistics for the ACAD-TRAIN training set and the ACAD-BENCH benchmark, including number of instances, number of translation directions, and mean (± standard deviation) source (Src) and target (Tgt) paragraph lengths computed using length in characters.

2. Related work

In the field of Machine Translation, large-scale multilingual corpora, often composed of synthetic or web-mined data, predominate. One of the most widely used repositories for MT is OPUS (Tiedemann, 2012), which aggregates nearly all publicly available parallel resources and provides a comprehensive overview of available parallel data. Its collection includes massive multilingual corpora such as CCMatrix (Schwenk et al., 2020), NLLB (Team et al., 2022), HPLT (de Gibert et al., 2024), or NTEU (García-Martínez et al., 2021).

As pointed out by Kreutzer et al. (2022) and Ranathunga et al. (2024), this web-mined data is generally characterized by issues such as poor alignment, reduced accuracy, high levels of noise, language and domain mismatch, and it typically requires extensive filtering and preprocessing before being suitable for training MT models (Steingrímsson et al., 2023).

In addition to generic corpora, a number of domain-specific datasets have been developed to support the fine-tuning of MT models on specialized domains. These resources are typically carefully filtered and curated, resulting in higher-quality data. Within the scientific domain, several examples exist. The Scielo corpus (Neves et al., 2016) is a parallel dataset of scientific publications for the biomedical domain in three language pairs: English to Spanish, English to French and English to Portuguese. The data was extracted from Scielo[4], a database of open access scientific publications with a focus on developing and emerging countries. In a related effort, Soares et al. (2018a) compiled a further version of the Scielo dataset by including parallel data for the directions English to Spanish, Portuguese to Spanish, and English to Portuguese.

Further examples are the Asian Scientific Paper Excerpt Corpus ASPEC (Nakazawa et al., 2016), a large-size parallel corpus of scientific paper abstracts in Japanese to English and Chinese to Japanese, and the CAPES TDC dataset (Soares et al., 2018b), a parallel corpus of theses and dissertations abstracts in English and Portuguese collected from the Brazilian CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior)[5] website. ed it, you can disable auto-conversion:

All these resources have a limited scope, since they are focused on a single domain (e.g., biomedicine) or a reduced number of languages. More recently, Roussis et al. (2022) introduced SciPar, a multilingual parallel corpus of theses and dissertations abstracts with 9.17 million sentence pairs in 31 language pairs including data extracted from 86 repositories and archives with openly available metadata. SciPar provides coverage across a broad range of scientific disciplines and serves as a key reference and inspiration for the present work. However, this prior work does not include any experimental applications or evaluations demonstrating the usefulness of the corpus for the training or assessment of MT systems, and it is limited by a non-commercial license. Building on this work, we introduce a new high-quality academic parallel corpus sourced from public institutions and demon-

[2] https://huggingface.co/BSC-LT/salamandraTA-2B-academic

[3] https://huggingface.co/BSC-LT/salamandraTA-7B-academic

[4] https://scielo.org

[5] https://www.gov.br/capes/pt-br

strate its practical value for fine-tuning. We include translation directions that were not covered by previous works. Additionally, we create a manually curated test set covering the most frequent language pairs in our training data, designed to support reliable evaluation of MT performance in the academic domain. Importantly, we release both portions of the dataset and the fine-tuned models under a permissive CC BY 4.0 license, providing a broadly accessible new resource and promoting its unrestricted use across the scientific community.

3. Methodology

This section describes the end-to-end procedure used to construct our dataset, from initial harvesting through final preprocessing.

3.1. Abstract Pair Harvesting

The core of our dataset consists of parallel paragraphs corresponding to abstracts, which are extracted from a variety of scientific, academic, and governmental repositories (most of them coming from Spanish institutions). All data are harvested via the OAI-PMH [6] (Open Archives Initiative - Protocol for Metadata Harvesting) interface. It is important to note that the harvested material does not include the full texts of the scientific works, which may be subject to more restrictive licenses, but consists exclusively of their abstracts. These abstracts have been obtained from metadata that institutions have deliberately made publicly available via the OAI-PMH protocol. No web scraping has been performed. According to the specific policy of the OAI-PMH repository, these abstracts are made available under the terms of the Creative Commons CC0 1.0 Universal license [7]. A complete list of the institutions and their repository URLs is provided in Appendix A.

During harvesting, we process each metadata record's `<description>` field. Whenever a record contains multiple description elements, we use the LaBSE sentence-transformer model (Feng et al., 2022) to embed each element into a shared multilingual latent space and compute pairwise cosine similarities. Since LaBSE is limited to 512-token sequences, we apply the SLIDE sliding window approach (Raunak et al., 2024) to process longer inputs. Pairs of descriptions are considered valid candidate translations if their cosine similarity is greater than 0.8.

To filter out off-target translations, we apply GlotLID (Kargaran et al., 2023) for language identification, discarding any pair in which either paragraph has a probability language score below 0.8. Then, we restrict our dataset to the most represented languages in the source repositories: English, Spanish, French, Catalan, Portuguese, German, Italian, Galician, Basque, Dutch, Greek and Asturian. Finally, we discard any pair in which either paragraph contains fewer than 40 characters, in order to reduce noise from overly brief instances. Following these filtering steps, we obtain 855,874 parallel paragraphs.

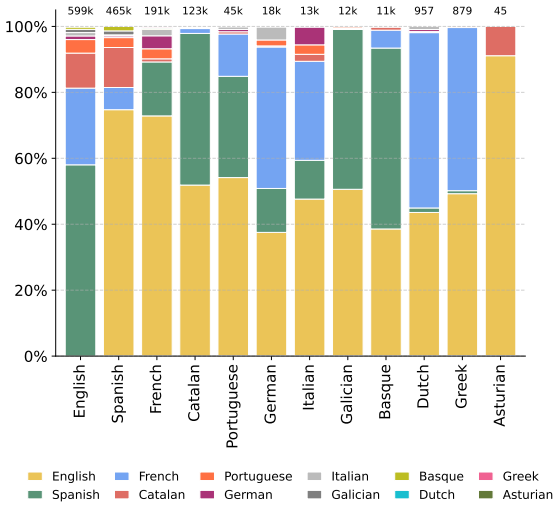

Figure 2: Relative distribution of data across languages in ACADATA. Each bar represents a source language, with segments showing the proportion of sentence pairs per target language. Percentages are normalized per language, while absolute pair counts are indicated above each bar.

3.2. Normalization and deduplication

After collecting all translation pairs, we apply a cleaning and normalization pipeline followed by deduplication to improve consistency and remove redundant entries.

First, any leading language markers (e.g., "(Spanish)", "[eng]") are stripped from each segment. We then normalize punctuation and typography by converting all variants of quotation marks and apostrophes to their ASCII equivalents, replacing masculine ordinals ("º") with degree symbols ("°"), and converting any superscript or subscript digits to regular digits. Next, we remove common inline markers (short bracketed or parenthesized codes, leading "//" or ":"), collapse simple HTML tags (e.g.,
, <i>, <b>), and collapse multiple whitespace characters into single spaces.

After normalization, we deduplicate at the pair level by constructing a key from the first 300 characters of each source-target segment and dropping any duplicate keys. This process yields a final set

[6] https://www.openarchives.org/pmh/

[7] https://oai-openedition.readthedocs.io/en/latest/license.html

of 733,709 clean, unique translation pairs. We use both directions for each pair, meaning the total number of translation instances is 1,467,418, ready for splitting into training and benchmarking subsets. The relative distribution of translation pairs across languages is shown in Figure 2, with absolute counts indicated above each bar. An exhaustive analysis is provided in Appendix B.

3.3. Benchmarking Set Splitting & Curation

After constructing the full dataset, we partition it into a large training set and a smaller, manually curated set. The latter contains a total of 2,972 pairs, which we also use in both directions, summing up to 5,944 translation instances. These instances are chosen as follows:

- **Language coverage** The test set includes the six most represented language pairs from the training data (see Table 2).
- **Size distribution** We sample instances from the eligible pairs so that the distribution of language pairs in the test set closely reflects that of the full training set. We use the number of translation pairs in each language pair as a weight to obtain the benchmarking split.

A detailed description of the construction process for ACAD-BENCH is provided in Appendix C.

After sampling, we manually curate the test set. We use the previously computed cosine similarities as a heuristic to identify possible low-quality translation pairs. Any pair with a similarity score below a threshold of 0.91, is post-edited by native speakers in ACAD-BENCH covered languages[8].

4. ACADATA

This section presents the ACADATA dataset, providing an analysis of its training and benchmarking splits. Throughout the tables, we report statistics for only one translation direction per language pair, since cosine similarity scores are symmetrical and source/target character counts simply swap in the reverse direction.

4.1. Training split: ACAD-TRAIN

The training set consists of 1,467,418 translation instances. Beyond applying embedding-based similarity filters, language-identification filters, normalization, and deduplication, no further processing is performed. We cover 96 translation directions. Table 2 gives an overview of the six most frequent language pairs (accounting for 96.5% of the ACADATA training set); all other directions are grouped under "Other". The last column of the table reports the average cosine similarity between LaBSE embeddings of the paired paragraphs. The relatively high scores reflect the high quality of the translation pairs. Full statistics are reported in Appendix B.

Lang Pair	Count	Src μ	Tgt μ	Cos μ
en–es	380,205	1,080	1,184	0.93
en–fr	141,972	900	979	0.90
en–ca	70,760	1,380	1,288	0.92
ca–es	62,264	1,176	1,204	0.98
es–fr	32,575	915	918	0.95
en–pt	25,916	1,010	1,042	0.90
Other	25,519	1,050	1,018	0.93
Overall	**739,211**	**1,070**	**1,126**	**0.93**

Table 2: Summary of the six largest language-pair subsets (all others are aggregated under "Other") in ACAD-TRAIN. Mean paragraph length (measured in characters) and mean cosine similarity. Counts are shown for one direction only (i.e. half of the total bidirectional instances).

4.2. Benchmarking split: ACAD-BENCH

The benchmarking set comprises the most common language pairs, sampled to preserve a similar size distribution, and contains 5,944 instances in total. Table 3 summarizes its statistics. Again, a detailed analysis is provided in Appendix B.

Lang Pair	Count	Src μ	Tgt μ	Cos μ
en–es	2,161	1,102	1,203	0.93
en–fr	333	897	969	0.91
en-ca	210	1,190	1,115	0.92
ca–es	188	1,290	1,317	0.98
es–fr	46	798	795	0.95
en–pt	34	1,037	1,065	0.90
Overall	**2,972**	**1,091**	**1,169**	**0.93**

Table 3: Summary of ACAD-BENCH. Mean paragraph length (measured in characters) and mean cosine similarity. Counts are shown for one direction only (i.e. half of the total bidirectional instances).

To understand the domain coverage of ACAD-BENCH, we classify its instances into 26 domains using NVIDIA's multilingual-domain classifier[9]. The 15 most frequent domains account for 94.6% of the

[8] In total, 114 instances were post-edited.

[9] https://huggingface.co/nvidia/multilingual-domain-classifier

2,972 instances[10]. The four largest domains: People and Society (806 instances, ≈21%), Health (376, ≈9.8%), Jobs and Education (325, ≈8.5%), and Science (216, ≈5.7%), together account for over 45% of all examples. A second group of domains: Arts and Entertainment (209), Computers and Electronics (189), Books and Literature (149), and Business and Industrial (130), each represent between 3 and 5% of the benchmark. The remaining 17 domains span a variety of topics: from Law and Government (102) and News (90) down to the smallest categories such as Real Estate (6), Shopping (5), Adult (4), and Online Communities (2), and make up the final ≈30% of instances. This diversity ensures that ACAD-bench covers both widely studied areas and more specialized subjects. Appendix D provides the list with the 26 classes identified by the classifier, a description of our methodology, and the complete distribution.

5. Experiments

To assess the usefulness of the ACADATA dataset, we fine-tune two LLMs on its training split and evaluate them on the benchmarking split. We then compare their performance against strong baselines. Previous work on adapting LLMs for long-context machine translation typically follow a two-stage training pipeline. In the first stage, models are fine-tuned on sentence-level parallel data. The second stage then adapts the model to handle longer contexts by training it on long-context parallel data (Zhang et al., 2018; Wu et al., 2024). Following Ramos et al. (2025), we skip the initial stage and continue supervised fine-tuning on ACADATA dataset by using two instructed models from the SALAMANDRA family of LLMs (Gonzalez-Agirre et al., 2025) that have already been trained on sentence-level parallel data in the instruction-tuning stage. These models were pre-trained from scratch on highly multilingual data and then instruction-tuned to improve performance on all the languages covered in ACADATA dataset. Specifically, we experiment with the instructed versions of SALAMANDRA-2B[11] and SALAMANDRA-7B[12] LLMs.

5.1. Formatting

We format each instruction using the commonly adopted `chatml` template (OpenAI, 2023) for instruction tuning.

5.2. Implementation details

We fine-tune SALAMANDRA LLMs using the `FastChat` (Zheng et al., 2023) and `DeepSpeed` (Rasley et al., 2020) frameworks on 32 NVIDIA H100 GPUs.

We train the models for one epoch with a per-GPU batch size of 1 and 16 gradient-accumulation steps, resulting in an effective batch size of 512. The learning rate was linearly warmed up over the first 85 steps, reaching a peak of 1×10^{-5}, and then decayed using a cosine schedule. SALAMANDRA models were trained on a context length of 8,192 tokens, which was sufficient for our training data. Thus, we kept the original maximum sequence length.[13]

5.3. Evaluation

Datasets We conduct experiments on two datasets. Our primary evaluation is on our ACAD-BENCH dataset, which targets the academic domain. We also use the general-domain WMT24++ test set (Deutsch et al., 2025) for two purposes: (i) to assess long-context improvements on the same English-centric language pairs as ACAD-BENCH, and (ii) to verify that fine-tuning on ACAD-TRAIN does not degrade performance on general-domain text, using the language directions common to both ACAD-TRAIN and WMT24++.

Baselines We compare the fine-tuned models with three categories of Machine Translation systems on ACAD-BENCH: ■ dedicated Massively Multilingual Neural Machine Translation (MMNMT) models, ■ general purpose open-weights LLMs and ■ large-scale proprietary LLMs.

- ■ **Dedicated MMNMT models:** we report MADLAD400-7B (Kudugunta et al., 2023): A widely-used encoder-decoder model supporting more than 400 languages that has been trained with sentence-level parallel data.
- ■ **General purpose open-weights LLMs:** we use LLAMA3-8B (Grattafiori et al., 2024) and GEMMA3-27B (Team et al., 2025) models.
- ■ **Large-scale proprietary LLMs:** we evaluate GPT-4.1-MINI, GPT-4.1-NANO (Achiam et al., 2023), GEMINI-2.0-FLASH and GEMINI-2.5-FLASH (Team et al., 2023).

Inference For the fine-tuned models, LLAMA3-8B and MADLAD400-7B we perform inference locally using beam search decoding with a beam size of 5. For large-scale proprietary LLMs, we access the

[10] For each translation pair, we employ only one direction, resulting in a total of 2,972 classified instances (5,944 / 2).

[11] https://huggingface.co/BSC-LT/salamandra-2b-instruct

[12] https://huggingface.co/BSC-LT/salamandra-7b-instruct

[13] With this configuration, the 2B model required $\approx 16h$ and the 7B model $\approx 39h$ to complete fine-tuning.

Direction	Model	d-BLEU	BP	BLONDE	COMET	COMET-KIWI
xx→en	GPT-4.1-MINI	46.03	**1.00**	0.60	**0.84**	0.77
	GPT-4.1-NANO	41.30	0.97	0.55	**0.84**	**0.78**
	GEMINI-2.0-FLASH	48.65	**1.00**	0.61	**0.84**	0.77
	GEMINI-2.5-FLASH	45.10	0.98	0.58	**0.84**	0.77
	LLAMA3-8B†	43.12	0.99	0.56	0.83	0.76
	GEMMA3-27B†	46.37	0.98	0.59	**0.84**	0.77
	MADLAD400-7B†	38.69	0.86	0.51	0.81	0.77
	SALAMANDRA-2B†	37.09	0.92	0.52	0.82	0.75
	+ ACAD-TRAIN	48.45	**1.00**	0.61	0.83	0.76
	SALAMANDRA-7B†	45.87	0.99	0.59	0.83	0.76
	+ ACAD-TRAIN	**50.07**	**1.00**	**0.62**	**0.84**	0.76
en→xx	GPT-4.1-MINI	45.01	0.99	-	0.86	**0.82**
	GPT-4.1-NANO	43.78	**1.00**	-	0.86	**0.82**
	GEMINI-2.0-FLASH	48.00	0.99	-	**0.87**	**0.82**
	GEMINI-2.5-FLASH	47.75	0.99	-	**0.87**	**0.82**
	LLAMA3-8B†	39.87	0.99	-	0.85	0.81
	GEMMA3-27B†	46.29	0.99	-	0.86	**0.82**
	MADLAD400-7B†	36.08	0.82	-	0.83	0.80
	SALAMANDRA-2B†	32.91	0.90	-	0.83	0.78
	+ ACAD-TRAIN	46.86	0.98	-	0.86	0.81
	SALAMANDRA-7B†	42.55	0.98	-	0.86	0.81
	+ ACAD-TRAIN	**49.20**	0.98	-	0.86	0.81
xx→es	GPT-4.1-MINI	60.60	0.98	-	0.86	**0.82**
	GPT-4.1-NANO	57.88	**0.99**	-	0.86	**0.82**
	GEMINI-2.0-FLASH	62.02	0.99	-	0.86	**0.82**
	GEMINI-2.5-FLASH	61.43	0.98	-	**0.87**	**0.82**
	LLAMA3-8B†	55.4	0.98	-	0.86	0.81
	GEMMA3-27B†	60.71	0.98	-	0.86	**0.82**
	MADLAD400-7B†	43.44	0.76	-	0.83	0.81
	SALAMANDRA-2B†	50.09	0.92	-	0.85	0.80
	+ ACAD-TRAIN	61.97	0.98	-	0.86	**0.82**
	SALAMANDRA-7B†	57.55	0.98	-	0.86	**0.82**
	+ ACAD-TRAIN	**63.60**	0.98	-	0.86	**0.82**
es→xx	GPT-4.1-MINI	54.19	**0.99**	-	**0.86**	**0.81**
	GPT-4.1-NANO	51.95	**0.99**	-	**0.86**	**0.81**
	GEMINI-2.0-FLASH	60.28	**0.99**	-	**0.86**	**0.81**
	GEMINI-2.5-FLASH	57.61	**0.99**	-	**0.86**	**0.81**
	LLAMA3-8B†	52.12	**0.99**	-	0.85	0.80
	GEMMA3-27B†	57.31	**0.99**	-	**0.86**	**0.81**
	MADLAD400-7B†	40.13	0.79	-	0.83	**0.81**
	SALAMANDRA-2B†	47.84	0.94	-	0.84	0.80
	+ ACAD-TRAIN	60.09	**0.99**	-	**0.86**	**0.81**
	SALAMANDRA-7B†	55.65	0.98	-	**0.86**	0.80
	+ ACAD-TRAIN	**61.61**	**0.99**	-	**0.86**	**0.81**

Table 4: Aggregated results for the xx↔en and xx↔es translation directions in ACAD-BENCH dataset. Baselines are grouped into **dedicated MMNMT models**, **medium- to small-sized open-weights models** and **large-scale proprietary general models**. Models with open weights are marked with †. For every metric, the top-scoring system is shown in bold, while the next two best systems for each direction are underlined.

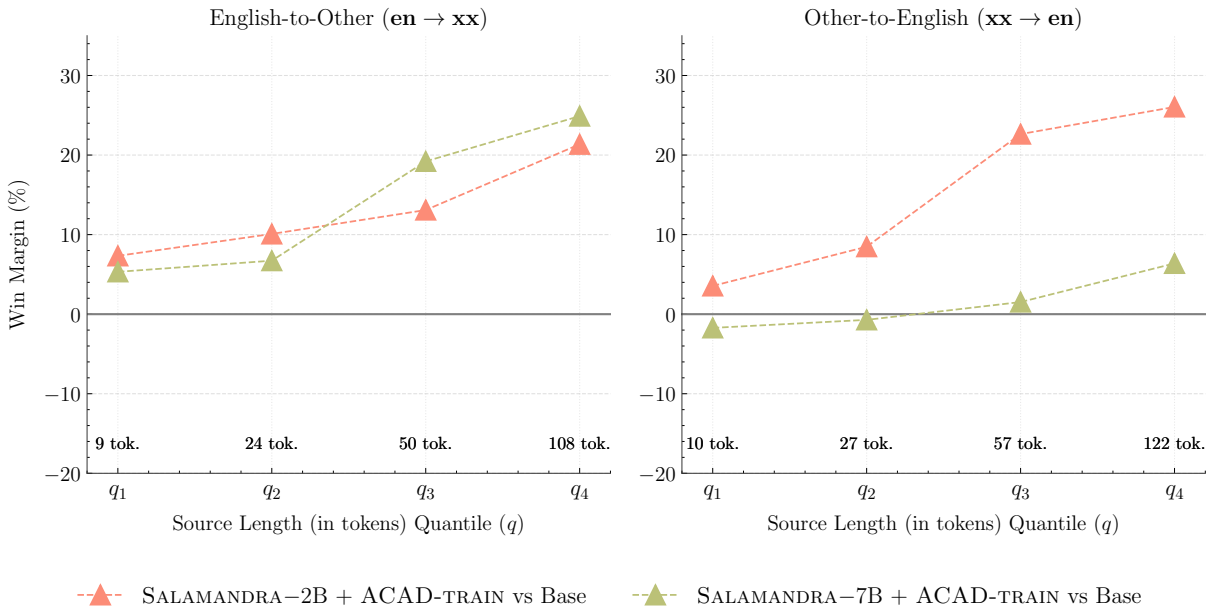

Figure 3: Win margin (%) over the base model as a function of source length, divided into quartiles (q_1–q_4) based on token count. We compare the performance of Salamandra-2B + ACAD-train and Salamandra-7B + ACAD-train against the base model across translation directions in the WMT24++ benchmark. The left plot shows results for English-to-Other (en→xx), and the right for Other-to-English (xx→en) directions.

models via their respective closed APIs, using their default generation settings. Finally, for evaluating the large-scale open-weight Gemma3-27b model we use Google API. More details about inference interfaces can be found in Appendix E.

Metrics System performance is evaluated using several metrics targeting different aspects of translation quality. For all translation directions, we report document-level BLEU (d-BLEU[14]) (Papineni et al., 2002) with its brevity penalty (BP).

We additionally evaluate translation quality using two learned regression-based metrics: COMET[15] (Rei et al., 2022a) and COMET-Kiwi[16] (Rei et al., 2022b). Learned regression-based metrics have been proved useful to evaluate translation quality at the paragraph level (Deutsch et al., 2023).

To evaluate discourse-level phenomena, we use Blonde (Jiang et al., 2022), a metric designed to capture discourse coherence through a set of automatically extracted features. Following Vernikos et al. (2022), we only evaluate Blonde when translating into English, since this metric depends on entity taggers and discourse markers trained exclusively in English.

Finally, to evaluate long-context improvements, we report the win margin based on BLEU on the WMT24++ dataset for the English-centric directions in ACAD-bench across different source sentence lengths. Details on how win margin is computed can be found in Appendix F.

Additionally, we report BLEU and COMET scores for the fine-tuned models for each of the 62 directions shared between WMT24++ and ACAD-train and analyze the overall performance in the following section.

5.4. Results

Table 4 presents the performance of all evaluated systems on ACAD-bench. We aggregate results for all models considering xx→en, en→xx, xx→es and es→xx directions. We find that fine-tuning on ACAD-train improves the 7B and 2B fine-tuned models by an average of +6.08 and +12.36 d-BLEU points over the base model, respectively. The fine-tuned models consistently outperform all other evaluated models, including large-scale proprietary systems in d-BLEU metric. We also report MadLad400-7b, a MMNMT model in Table 4. One of the reasons why MadLad400-7b may be performing worse is that it has been trained on sentence-level corpora, while ACAD-bench evaluates paragraph-level translation.

Notably, Salamandra-7B + ACAD-train is the top-performing model, achieving the highest d-BLEU score across all translation directions. It outperforms the strongest proprietary baseline, Gemini-

[14] `Signature: nrefs:1| case:mixed| eff:no| tok:13a| smooth:exp| version:2.3.1`

[15] https://huggingface.co/Unbabel/wmt22-comet-da

[16] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

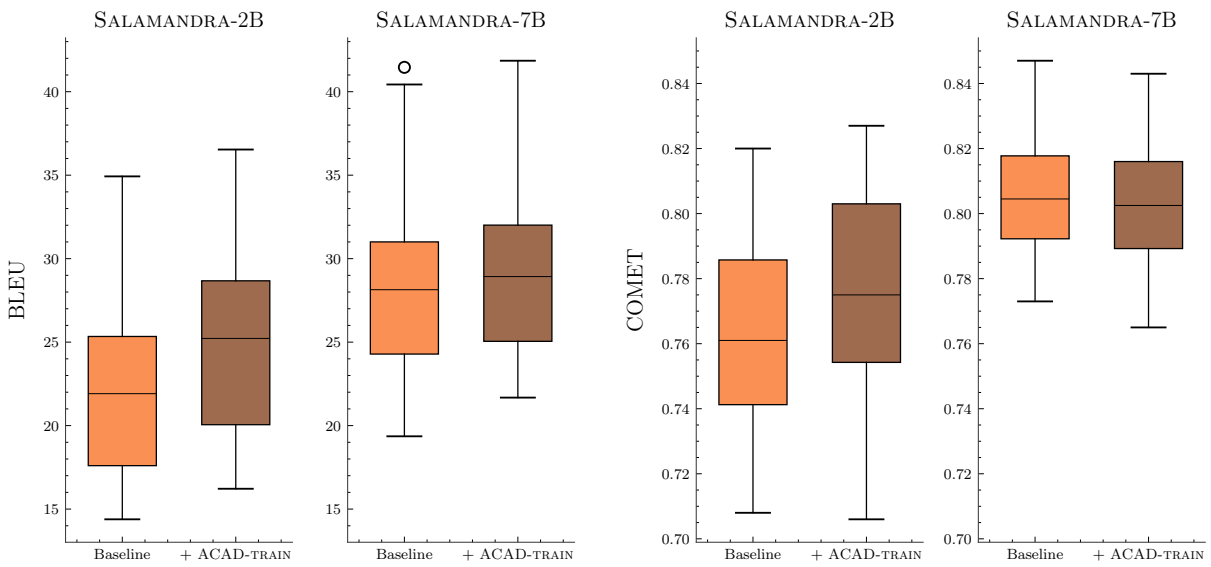

Figure 4: Boxplots for the 62 directions shared between WMT24++ and ACAD-TRAIN. The plots show the distribution of BLEU scores (left) and COMET scores (right). Each boxplot compares the model's performance on WMT24++ dataset before (Baseline) and after fine-tuning on the ACAD-TRAIN dataset (+ ACAD-TRAIN).

2.0-FLASH, by an average of 1.39 d-BLEU points. Even the smaller fine-tuned 2B model outperforms most proprietary models and significantly closes the gap with GEMINI-2.0-FLASH.

For BLONDE, which evaluates discourse and context understanding, the fine-tuned models improve by +0.03 and +0.09 points over the base 7B and 2B models respectively. This result demonstrates a stronger ability to handle context-dependent translation phenomena. However, on semantic metrics like COMET and COMET-KIWI, the fine-tuned models achieve similar performance as the other baselines, with proprietary models slightly leading in English-centric directions.

Improved quality in long-context translation Figure 3 reports win margin on the same English-centric language pairs as ACAD-BENCH, evaluated on WMT24++. Translation quality is divided into four quartiles by source sentence length (measured with the model's tokenizer). The performance gap between the fine-tuned models and their base counterparts grows as the length of the source sentence increases when translating out of English. In fact, when translating out of English, we observe an average win ratio improvement of 24.9% and 21.3% in the longest quartile (q4) for the 7B and 2B models respectively. On the contrary, when translating into English, the 2B model shows a strong 26% improvement in win margin, compared to the 7B model's average gain of 6.4% in the longest quartile. These results are consistent with the improvements in the brevity penalty (BP) reported for both the 7B and 2B models in Table 4, indicating that fine-tuned models are better able to maintain output length parity in long-context translations.

Improved BLEU performance across WMT24++ language directions Figure 4 shows the summary of results when fine-tuning with the ACAD-TRAIN dataset, evaluated on the WMT24++ test set. The results demonstrate a clear and positive impact on translation quality as measured by BLEU. The 2B model fine-tuning's average score increases by +2.64 BLEU. The larger 7B model also shows improvements, with its average score improving by +1.09 BLEU. In contrast, the impact on COMET scores, which measure semantic similarity, is smaller. The 2B model's COMET score improves by +0.01 points on average, while the 7B model sees a minor decrease of -0.01.

Overall, these findings indicate that adapting models to the academic domain leads to small improvements on lexical-based metrics, while maintaining comparable performance on semantic-based metrics in a general domain dataset. We hypothesize that fine-tuning with academic data encourages translations that are more technical and formal in style. As a result, improvements are greater on lexical-based metrics such as BLEU, while metrics that focus on semantic similarity, like COMET, show minor changes. We also find that smaller models tend to benefit more from fine-tuning on ACAD-TRAIN, whereas larger models show more stable performance across metrics in a general domain dataset, especially when translating into English.

6. Conclusions

In this work, we describe the creation process of the ACADATA dataset, a novel, high-quality parallel corpus of academic abstracts. Comprising approximately 1.5 million human-generated paragraph pairs (ACAD-TRAIN) and a manually curated evaluation set of nearly 6,000 translations (ACAD-BENCH), our dataset provides accurate, high-quality translations in the scientific domain. By fine-tuning two open-weight LLMs on ACAD-TRAIN, we demonstrate consistent gains in academic translation quality, up to +12.4 d-BLEU for the 2B parameter model and +6.1 d-BLEU for the 7B parameter model, outperforming proprietary models on ACAD-BENCH.

Beyond the academic domain, we also show that training with ACAD-TRAIN yields significant improvements in long-context translation on a standard general-domain benchmark: in the longest source-length quartile, the fine-tuned models achieve up to a 24.9% gain in win ratio when translating out of English.

By releasing ACAD-TRAIN, ACAD-BENCH, and the fine-tuned models under permissive licenses (CC BY 4.0 for the datasets and Apache 2.0 for the models), we offer the community a robust foundation training dataset and evaluation benchmark for advancing the development of machine translation systems in the academic domain.

Ultimately, with this work, we aim to help bridge communication across the global scientific community, and make research more discoverable and accessible regardless of the language it was originally published in.

Limitations

In this work, we explain how we built ACADATA and employed its benchmarking split to evaluate a set of models. However, the scope of the evaluated systems remains limited; future research could extend this analysis to a broader range of models. Additionally, since several baseline systems are proprietary, it is not possible to assess whether the data collected to create ACADATA was included in their training data, which could potentially influence the results.

Ethical Statement

This work focuses on describing the creation of a dataset for academic machine translation sourced from public data, and the fine-tuning of machine translation systems using its training set. The impact of fine-tuning on potential biases, such as gender bias, is left out of the scope of this work. All models and datasets used in our experiments are based on publicly available resources, that may contain inherent biases.

Acknowledgements

This work/research has been promoted and financed by the Government of Catalonia through the Aina project.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335, 2022/TL22/00215334

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project Desarrollo de Modelos ALIA.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Mikko Aulamo, Ona de Gibert, Sami Virpioja, and Jörg Tiedemann. 2023. Unsupervised feature

selection for effective parallel corpus filtering. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 31–38, Tampere, Finland. European Association for Machine Translation.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

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

Mercedes García-Martínez, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis. 2021. Neural translation for European Union (NTEU). In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 316–334, Virtual. Association for Machine Translation in the Americas.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Valle Ruíz-Fernández, and Marta Villegas. 2025. Salamandra technical report.

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

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. 2016. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

NVIDIA. n.d. Nemocurator multilingual domain classifier. `https://huggingface.co/nvidia/multilingual-domain-classifier`. Accessed: 2025-08-04.

OpenAI. 2023. ChatML.

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.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André FT Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. *arXiv preprint arXiv:2504.12140*.

Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

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. 2022a. 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, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. SciPar: A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France. European Language Resources Association.

Dimitris Roussis, Sokratis Sofianopoulos, and Stelios Piperidis. 2024. Enhancing scientific discourse: Machine translation for the scientific domain. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 275–285, Sheffield, UK. European Association for Machine Translation (EAMT).

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2020. Ccmatrix: Mining billions of high-quality parallel sentences on the web.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018a. A large parallel corpus of full-text scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Felipe Soares, Gabrielli Harumi Yamashita, and Michel Jose Anzanello. 2018b. A parallel corpus of theses and dissertations abstracts. In *International Conference on Computational Processing of the Portuguese Language*, pages 345–352. Springer.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Daniel Stockemer and Michael J. Wigginton. 2019. Publishing in english or another language: An inclusive study of scholar's language publication preferences in the natural, social and interdisciplinary sciences. *Scientometrics*, 118(2):645–652.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona

Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A. Data Sources

This appendix contains a complete list of the names and URLs of the source repositories from which the translation pairs are obtained, along with the number of pairs extracted from each source before deduplication. This information is shown in Table 5.

B. Dataset Analysis

In this appendix, we present a detailed analysis of the ACADATA dataset. For each language pair, we report the number of paragraph pairs, length statistics (in characters), and cosine similarity scores between source and target embeddings, which serve as a proxy for semantic relatedness. In Table 6, we show the mapping between the BCP-47 language code and the language name for the languages covered in ACAD-TRAIN.

Table 7 summarizes ACAD-TRAIN main statistics. All language pairs exhibit strong alignment, with all mean and median cosine similarities above 0.85, and most above 0.90, indicating that the translated paragraphs closely match the source semantics. Average source and target lengths hover around 1,000 characters, while large standard deviations reflect a wide variation in paragraph length, proving the diversity of the dataset.

Table 8 shows the summary of ACAD-BENCH. Every pair exceeds a mean similarity of 0.90 (overall mean ± std; median = 0.93 ± 0.03; 0.93), underscoring the high semantic fidelity of the test translations. As with the training data, paragraph lengths in the test set vary considerably.

For both cases, in order to reduce redundancy, only one direction is shown. Changing the translation direction would imply switching the source length and target length values, while the cosine similarity scores would remain the same.

C. ACAD-BENCH Extraction

We want to make sure that the benchmarking set reflects the same language-pair distribution as the full training data while remaining manageable in size. First, we identify all language pairs (unordered) in ACAD-TRAIN with at least 1,000 total pairs. Let P be this set of eligible language pairs. We then gather all candidate test-set paragraph pairs whose source–target languages belongs to P. To sample up to 3,000 test pairs, we draw without replacement in proportion to each pair's frequency in the training set, yielding a preliminary pool of $3{,}000$ distinct paragraph pairs. Next, we enforce a minimum-per-pair threshold: we discard any language pair for which fewer than 10 sentences have been sampled. This pruning leaves exactly the six most frequent language pairs from the training data. Finally, for

Repository	Pairs	Repository	Pairs	Repository	Pairs
al-qantara	442	estudios medievales	947	uca	4427
anales cervantinos	279	fs	661	ucam	1036
anuario musical	326	gredos	31665	ucasal	1
arbor	1265	helvia	7980	uclm	2502
archivo español de arte	528	hispania csic	750	ucm	7844
archivo español de arqueología	528	humanum	1056	udc	24273
arcimis	975	ifapa	346	udimundus	43
arias montano	10150	ignacio larramendi	566	udl	1
arqueología de la arquitectura	280	indteca	477	ufv	354
asclepio	791	irta	25	ugr	22448
auditio	72	isciii	38	uja	1057
bcnroc	3475	jardín botánico de madrid	458	uji	17223
biotecnia	345	maco	1228	ulerevistas	1487
brújula	133	manipulus	1	ulpgc	7198
cgate	240	o2	41462	uni peru	195
claves jurídicas	15	ojs	766	unia	633
collectanea botanica	239	open edition	244720	uniboyaca	360
cora	101	pse	87	unican	15338
csic	9588	produccion cientifica luz	88	uniovi	246
cuadernos de estudios gallegos	295	pubmed	22813	unir	5383
culture history d-journal	265	r-usj	517	unirioja	556
dau	2478	rccs	65	universidad peruana de ciencias	687
diffundit	766	rcg	409	unizar	226
digi uv	320	rcj	15	unla	599
digitium	23268	rdtecnocampus	59	upc	44275
docta	20990	redined	1250	upf	8269
e-buah	9094	rediumh	6354	upv/ehu	16360
e-cienciadatos	1028	sjar	902	urv	1230
e-ucjc	43	tdx	31858	usil	371
ebiltegia	3310	ua	25258	uv	9212
emd	617	uab	96620	uva	12002
emerita	337	uam	11066	uvic	3462
enfermería cuidándote	52	uasd	217	zenodo	386
estudios americanos	472	ub	19339	zoilomarinello	740
estudios geográficos	726	ubu	1554		

Table 5: Sources of translation pairs with paragraph counts before deduplication.

each selected sentence pair (s, t), we generate two directional translation instances $(s \rightarrow t)$ and $(t \rightarrow s)$, resulting in a total of $2 \times 2{,}972 = 5{,}944$ test instances.

Language Code	Language
ast	Asturian
ca	Catalan
de	German
el	Greek
es	Spanish
en	English
eu	Basque
fr	French
gl	Galician
it	Italian
nl	Dutch
pt	Portuguese

Table 6: Mapping from BCP-47 language codes to full language names.

D. ACAD-BENCH Domain Distribution

Table 9 provides descriptions of the domains predicted by NVIDIA's Multilingual Domain Classifier.

Figure 5 illustrates that ACAD-BENCH's content is concentrated in a handful of major areas. The four largest domains: People and Society (806 instances, ≈21%), Health (376, ≈9.8%), Jobs and Education (325, ≈8.5%), and Science (216, ≈5.7%), together account for just over 45% of all examples. A second group of domains: Arts and Entertainment (209), Computers and Electronics (189), Books and Literature (149), and Business and Industrial (130), each represent between 3 and 5% of the benchmark.

The remaining 17 domains span a variety of topics: from Law and Government (102) and News (90) down to the smallest categories such as Real Estate (6), Shopping (5), Adult (4), and Online Communities (2), and collectively make up the final ≈30% of instances. This breadth ensures that ACAD-BENCH covers both widely studied areas and more specialized subjects.

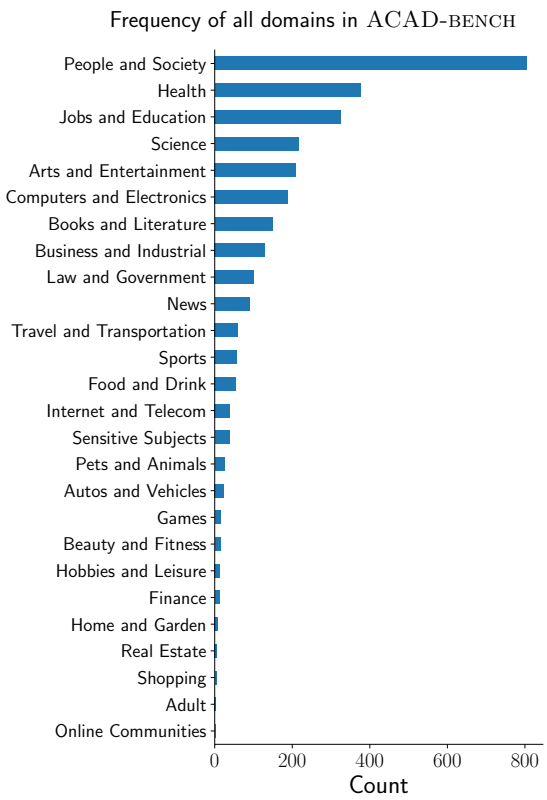

Figure 5: Bar plot showing all the domains in ACAD-BENCH.

E. Inference via API

This appendix section describes our end-to-end inference pipeline across different model families. We outline the prompt templates used, highlight challenges in obtaining strictly formatted outputs, and specify the interfaces and settings leveraged for each model, whether via remote APIs or local deployments.

E.1. Challenges with Model Output Formatting

We observed that certain conversational models frequently prepend or append unsolicited comments instead of returning a raw translation. Examples of such extraneous text include:

- "Here's the translation from {src} to {tgt}:"
- "Breakdown of choices: . . . "
- "Let me know if you'd like any specific parts clarified or alternative phrasing considered!"

We attribute this behavior to the conversational datasets used to train these models, which encourage them to add explanations, multiple options, or task descriptions, and make them unable to fully understand the specified format in the prompt. To enforce a strict "translation-only" output, we design an *adapted* prompt (Figure 6) that explicitly instructs the model to produce only the translated text, with no additional prefixes, suffixes, or comments.

E.2. API and Deployment Details

For each model family, we use the following inference interfaces and settings:

OpenAI API We access closed-source OpenAI models via the official OpenAI API[17] for the following models; GPT-4.1-MINI[18], GPT-4.1-NANO[19]. Since these endpoints do not accept custom generation hyperparameters, we rely on the default settings. Notably, both GPT-4.1-MINI and GPT-4.1-NANO follow the original "translation-only" instruction without requiring the adapted template.

Google API Inference on Google's GEMINI-2.5-FLASH[20], GEMINI-2.0-FLASH[21], and GEMMA3-27B is performed through the Google AI API[22]. These models likewise do not expose generation hyperparameters. All three models comply with the original prompt and do not require any template adaptations.

Local Hugging Face Deployments For open-source models, including SALAMANDRA (2B, 7B, and their fine-tuned variants) and LLAMA3-8B, we load model checkpoints in Hugging Face format and run inference locally. In these cases, we use beam search decoding, setting the beam size to 5.

F. Win Margin

To compare two systems, we define a *win* when the difference in BLEU scores between the alternative system and the baseline system exceeds a threshold of 5 points. Formally, let $s_i^{(A)}$ and $s_i^{(B)}$ denote the BLEU scores for source text i from the alternative system A and the baseline system B,

[17] https://openai.com/api/

[18] Retrieved July 21, 2025, via OpenAI API from https://platform.openai.com/docs/models/gpt-4.1-mini

[19] Retrieved July 21, 2025, via OpenAI API from https://platform.openai.com/docs/models/gpt-4.1-nano

[20] Retrieved July 24, 2025, via Google AI API from https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash

[21] Retrieved July 24, 2025, via Google AI API from https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

[22] https://ai.google.dev/

Prompt template used for inference

```
Only output the translation, only one option, no explanations or anything else:
Translate the following text from {src. language} to {tgt. language}.
{src. language}: {src. text}
{tgt. language}:
```

Figure 6: Adapted prompt for inference with models that cannot follow the original template format. Implemented to force a single, clean translation.

respectively. The *delta* is computed as:

$$\Delta_i = s_i^{(A)} - s_i^{(B)}.$$

We define a win as $\Delta_i > 1$, and a loss as $\Delta_i < -1$. Ties or negligible differences ($|\Delta_i| \leq 5$) are ignored. The **win rate margin** is then computed as:

$$\text{WinMargin} = 100 \cdot \frac{N_{\text{wins}} - N_{\text{losses}}}{N_{\text{valid}}}$$

where N_{wins} is the number of instances with $\Delta_i > 5$, N_{losses} is the number of instances with $\Delta_i < -5$, and $N_{\text{valid}} = N_{\text{wins}} + N_{\text{losses}}$ is the number of comparisons where a win or loss occurred.

To analyze performance with respect to length, we group the source text into Q quantiles based on their token length and report the win margin per quantile.

G. Results

G.1. Results test set

In Tables 10, 11, 12, and 13, we show detailed results per language pair for xx→en, en→xx, xx→es and es→xx directions, respectively, evaluated on ACAD-BENCH.

Pair	Count	Source Len. ($\mu \pm \sigma$; med)	Target Len. ($\mu \pm \sigma$; med)	Cosine Sim. ($\mu \pm \sigma$; med)
en-es	347479	1066±891; 896	1169±1007; 975	0.92±0.03; 0.93
en-fr	139350	899±420; 839	978±452; 915	0.90±0.03; 0.91
ca-en	63615	1318±1098; 1018	1230±1025; 958	0.92±0.03; 0.92
ca-es	56407	1136±973; 895	1162±992; 918	0.98±0.02; 0.98
es-fr	31301	915±448; 855	918±454; 856	0.95±0.02; 0.96
en-pt	24596	1005±432; 948	1037±449; 978	0.90±0.03; 0.90
es-pt	13952	1152±504; 1102	1096±484; 1043	0.97±0.02; 0.97
de-fr	7531	1017±508; 947	986±491; 922	0.92±0.03; 0.93
de-en	6587	1032±505; 965	927±460; 862	0.88±0.03; 0.88
es-eu	6272	780±725; 641	710±660; 581	0.92±0.03; 0.92
en-it	6128	910±496; 832	969±553; 884	0.92±0.03; 0.93
en-gl	5858	1314±700; 1236	1375±738; 1285	0.91±0.03; 0.91
fr-pt	5829	984±491; 907	922±458; 850	0.95±0.02; 0.96
es-gl	5618	1387±831; 1332	1319±795; 1271	0.98±0.01; 0.99
en-eu	4408	944±667; 774	938±678; 767	0.88±0.03; 0.88
fr-it	3863	951±428; 904	932±402; 886	0.95±0.02; 0.96
de-es	2339	992±467; 916	947±427; 872	0.93±0.03; 0.93
ca-fr	1806	915±489; 874	943±513; 903	0.96±0.02; 0.96
es-it	1513	1149±845; 966	1144±861; 957	0.96±0.02; 0.97
de-it	703	1124±419; 1082	1047±390; 1000	0.93±0.02; 0.94
eu-fr	621	499±387; 431	553±393; 486	0.89±0.04; 0.90
fr-nl	509	905±504; 892	909±503; 897	0.93±0.03; 0.94
el-fr	435	913±386; 912	900±383; 890	0.88±0.04; 0.89
el-en	433	916±399; 912	850±371; 843	0.87±0.03; 0.87
en-nl	417	1000±533; 977	1089±613; 1072	0.90±0.03; 0.91
it-pt	369	952±282; 954	891±271; 881	0.96±0.02; 0.96
de-pt	316	1040±342; 1038	892±311; 887	0.94±0.02; 0.94
ca-pt	308	1059±589; 910	1027±562; 898	0.96±0.03; 0.96
ca-it	273	1000±851; 822	1014±881; 799	0.96±0.02; 0.97
ca-eu	112	884±696; 743	848±627; 708	0.93±0.02; 0.93
ca-de	59	1122±953; 904	1228±1139; 976	0.93±0.03; 0.94
gl-pt	47	1811±1195; 1446	1755±1161; 1418	0.98±0.02; 0.98
ast-en	41	636±510; 573	633±527; 502	0.88±0.04; 0.89
ca-gl	33	1307±972; 1032	1282±911; 995	0.96±0.03; 0.96
es-nl	13	2887±1926; 2651	3045±2282; 1747	0.93±0.02; 0.94
fr-gl	13	2759±3853; 1156	2652±3638; 1106	0.94±0.04; 0.94
de-eu	10	1048±678; 873	951±559; 905	0.89±0.02; 0.89
it-nl	9	1048±894; 1130	1074±896; 1157	0.91±0.05; 0.92
el-es	8	2081±1300; 1817	2128±1417; 1723	0.88±0.04; 0.89
de-nl	6	1497±939; 1538	1442±834; 1553	0.91±0.04; 0.92
gl-it	5	870±342; 878	958±452; 898	0.95±0.02; 0.96
eu-pt	4	663±235; 629	631±238; 563	0.92±0.04; 0.94
ast-ca	4	67±15; 69	65±15; 63	0.98±0.01; 0.99
ca-el	3	1959±1181; 1287	2020±1111; 1470	0.87±0.05; 0.86
ca-nl	3	2997±1075; 3260	3293±1452; 3238	0.93±0.03; 0.93
de-gl	2	2668±2598; 2668	2498±2487; 2497	0.91±0.00; 0.91
eu-it	2	473±172; 472	484±74; 484	0.86±0.01; 0.86
eu-gl	1	655±–; 655	662±–; 662	0.95±–; 0.95

Table 7: Main statistics of the ACAD-TRAIN set, shown only for one direction.

Pair	Count	Src Len. ($\mu \pm \sigma$; med)	Tgt Len. ($\mu \pm \sigma$; med)	Cosine Sim. ($\mu \pm \sigma$; med)
en–es	2161	1102±809; 909	1203±879; 993	0.93±0.03; 0.93
en–fr	333	897±392; 836	969±429; 902	0.91±0.03; 0.92
ca–en	210	1190±949; 966	1115±904; 876	0.92±0.03; 0.92
ca–es	188	1290±1123; 915	1317±1132; 945	0.98±0.02; 0.98
es–fr	46	798±247; 775	795±244; 780	0.95±0.03; 0.95
en–pt	34	1037±408; 970	1065±416; 1036	0.90±0.03; 0.90
Overall	**2972**	**1091±779; 904**	**1169±832; 969**	**0.93±0.03; 0.93**

Table 8: Main statistics of the ACAD-BENCH set, shown only for one direction.

Domain Class	Description
Adult	Sexual content, pornography, or age-restricted material
Arts_and_Entertainment	Music, movies, theater, celebrities, pop culture
Autos_and_Vehicles	Cars, motorbikes, vehicle news and reviews
Beauty_and_Fitness	Skincare, cosmetics, wellness, workout routines
Books_and_Literature	Novels, literary criticism, poetry, book reviews
Business_and_Industrial	Enterprise, corporate, manufacturing, B2B topics
Computers_and_Electronics	Hardware, software, tech news, consumer gadgets
Finance	Banking, investing, personal finance, stock markets
Food_and_Drink	Recipes, restaurants, food culture, drinks
Games	Video games, board games, eSports, gaming culture
Health	Medical topics, mental health, wellness, diseases
Hobbies_and_Leisure	DIY, crafts, hobbies, leisure activities
Home_and_Garden	Home improvement, gardening, decor
Internet_and_Telecom	ISPs, web platforms, telecommunications
Jobs_and_Education	Career guidance, job listings, academic topics
Law_and_Government	Legislation, public policy, political topics
News	Journalism, current events, news reporting
Online_Communities	Forums, social platforms, user communities
People_and_Society	Culture, social issues, demographics
Pets_and_Animals	Pet care, wildlife, zoology topics
Real_Estate	Property listings, housing market, realty advice
Science	Research, scientific articles, STEM topics
Sensitive_Subjects	Controversial or delicate content (e.g. abuse, violence)
Shopping	E-commerce, product reviews, retail
Sports	Athletic events, scores, sports commentary
Travel_and_Transportation	Tourism, transit, travel guides

Table 9: Domain class descriptions for NVIDIA's multilingual domain classifier, based on manual inspection of sample instances.

Pair	Model	d-BLEU	BP	BLONDE	COMET	COMET-KIWI
es→en	GPT-4.1-MINI	48.06	1.00	0.60	0.84	0.78
	GPT-4.1-NANO	42.81	0.97	0.56	0.84	0.78
	GEMINI-2.0-FLASH	51.27	1.0	0.62	0.84	0.77
	GEMINI-2.5-FLASH	46.99	0.98	0.6	0.84	0.77
	LLAMA3-8B†	45.12	0.99	0.58	0.84	0.77
	GEMMA3-27B†	48.83	0.99	0.61	0.84	0.77
	MADLAD400-7B†	35.83	0.8	0.48	0.81	0.78
	SALAMANDRA-2B†	37.94	0.92	0.52	0.82	0.76
	+ ACAD-TRAIN	51.28	1.00	0.62	0.84	0.76
	SALAMANDRA-7B†	47.72	0.99	0.59	0.84	0.76
	+ ACAD-TRAIN	53.06	1.00	0.63	0.84	0.76
pt→en	GPT-4.1-MINI	48.44	0.99	0.62	0.84	0.75
	GPT-4.1-NANO	44.10	0.97	0.56	0.84	0.76
	GEMINI-2.0-FLASH	51.09	1.0	0.63	0.83	0.75
	GEMINI-2.5-FLASH	47.81	0.98	0.59	0.84	0.75
	LLAMA3-8B†	46.04	0.98	0.56	0.83	0.74
	GEMMA3-27B†	48.49	0.97	0.59	0.84	0.75
	MADLAD400-7B†	49.93	0.96	0.61	0.82	0.75
	SALAMANDRA-2B†	40.92	0.93	0.54	0.81	0.73
	+ ACAD-TRAIN	51.29	1.00	0.63	0.83	0.74
	SALAMANDRA-7B†	49.62	1.00	0.63	0.83	0.75
	+ ACAD-TRAIN	53.14	1.00	0.65	0.83	0.74
fr→en	GPT-4.1-MINI	38.76	0.99	0.54	0.83	0.79
	GPT-4.1-NANO	35.11	0.97	0.51	0.83	0.79
	GEMINI-2.0-FLASH	41.15	1.0	0.56	0.83	0.79
	GEMINI-2.5-FLASH	37.92	0.97	0.53	0.83	0.79
	LLAMA3-8B†	36.48	1.0	0.51	0.82	0.77
	GEMMA3-27B†	38.63	0.98	0.54	0.83	0.79
	MADLAD400-7B†	36.87	0.93	0.5	0.81	0.79
	SALAMANDRA-2B†	32.30	0.93	0.48	0.81	0.77
	+ ACAD-TRAIN	41.66	0.99	0.56	0.82	0.78
	SALAMANDRA-7B†	38.78	0.98	0.54	0.82	0.78
	+ ACAD-TRAIN	40.61	1.00	0.55	0.82	0.77
ca→en	GPT-4.1-MINI	48.85	1.00	0.62	0.85	0.76
	GPT-4.1-NANO	43.20	0.97	0.58	0.85	0.77
	GEMINI-2.0-FLASH	51.10	1.0	0.64	0.85	0.76
	GEMINI-2.5-FLASH	47.67	0.98	0.61	0.85	0.77
	LLAMA3-8B†	44.86	1.0	0.58	0.84	0.76
	GEMMA3-27B†	49.54	0.99	0.62	0.85	0.76
	MADLAD400-7B†	32.12	0.74	0.45	0.8	0.77
	SALAMANDRA-2B†	37.21	0.92	0.52	0.83	0.75
	+ ACAD-TRAIN	50.63	1.0	0.63	0.84	0.75
	SALAMANDRA-7B†	47.36	1.0	0.6	0.84	0.75
	+ ACAD-TRAIN	52.42	1.0	0.64	0.85	0.75

Table 10: Translation results for the xx→en language pairs in ACAD-BENCH dataset. Baselines are grouped into **dedicated MMNMT models**, **medium- to small-sized open-weights models** and **large-scale proprietary general models**. Models with open weights are marked with †.

Pair	Model	d-BLEU	BP	COMET	COMET-KIWI
en→es	GPT-4.1-MINI	50.60	0.98	0.86	0.83
	GPT-4.1-NANO	48.83	0.99	0.86	0.82
	GEMINI-2.0-FLASH	52.48	0.99	0.86	0.82
	GEMINI-2.5-FLASH	52.02	0.98	0.87	0.82
	LLAMA3-8B†	44.76	0.98	0.85	0.81
	GEMMA3-27B†	50.74	0.98	0.86	0.82
	MADLAD400-7B†	34.88	0.73	0.82	0.8
	SALAMANDRA-2B†	34.47	0.84	0.83	0.79
	+ ACAD-TRAIN	51.31	0.97	0.86	0.81
	SALAMANDRA-7B†	46.42	0.96	0.86	0.81
	+ ACAD-TRAIN	53.47	0.97	0.86	0.81
en→pt	GPT-4.1-MINI	47.38	1.0	0.86	0.81
	GPT-4.1-NANO	45.23	1.0	0.86	0.82
	GEMINI-2.0-FLASH	48.18	1.0	0.87	0.81
	GEMINI-2.5-FLASH	48.29	1.0	0.86	0.81
	LLAMA3-8B†	41.81	0.99	0.85	0.8
	GEMMA3-27B†	46.96	1.0	0.86	0.81
	MADLAD400-7B†	49.58	0.97	0.86	0.8
	SALAMANDRA-2B†	37.12	0.96	0.84	0.77
	+ ACAD-TRAIN	48.18	0.99	0.86	0.81
	SALAMANDRA-7B†	43.73	1.0	0.85	0.81
	+ ACAD-TRAIN	50.26	0.99	0.86	0.81
en→fr	GPT-4.1-MINI	37.55	1.0	0.86	0.84
	GPT-4.1-NANO	37.05	1.0	0.85	0.83
	GEMINI-2.0-FLASH	42.36	0.99	0.86	0.83
	GEMINI-2.5-FLASH	41.58	0.99	0.86	0.83
	LLAMA3-8B†	36.14	0.99	0.84	0.82
	GEMMA3-27B†	40.05	1.0	0.86	0.83
	MADLAD400-7B†	37.3	0.92	0.84	0.82
	SALAMANDRA-2B†	29.73	0.93	0.81	0.79
	+ ACAD-TRAIN	40.05	0.97	0.84	0.82
	SALAMANDRA-7B†	37.19	0.97	0.84	0.82
	+ ACAD-TRAIN	42.28	0.96	0.85	0.82
en→ca	GPT-4.1-MINI	44.51	0.98	0.88	0.82
	GPT-4.1-NANO	44.03	0.99	0.88	0.82
	GEMINI-2.0-FLASH	48.99	0.99	0.88	0.82
	GEMINI-2.5-FLASH	49.12	0.99	0.88	0.82
	LLAMA3-8B†	36.76	1.0	0.86	0.8
	GEMMA3-27B†	47.41	0.99	0.88	0.82
	MADLAD400-7B†	22.54	0.65	0.79	0.77
	SALAMANDRA-2B†	30.33	0.87	0.85	0.78
	+ ACAD-TRAIN	47.90	0.97	0.88	0.81
	SALAMANDRA-7B†	42.87	0.97	0.87	0.81
	+ ACAD-TRAIN	50.77	0.98	0.88	0.81

Table 11: Translation results for the en→xx language pairs in ACAD-BENCH dataset. Baselines are grouped into **dedicated MMNMT models**, **medium- to small-sized open-weights models** and **large-scale proprietary general models**. Models with open weights are marked with †.

Pair	Model	d-BLEU	BP	COMET	COMET-KIWI
fr→es	GPT-4.1-MINI	48.54	0.97	0.84	0.84
	GPT-4.1-NANO	46.74	0.97	0.84	0.84
	GEMINI-2.0-FLASH	48.78	0.98	0.84	0.84
	GEMINI-2.5-FLASH	48.97	0.97	0.84	0.84
	LLAMA3-8B†	44.26	0.97	0.83	0.83
	GEMMA3-27B†	48.66	0.97	0.84	0.84
	MADLAD400-7B†	50.84	0.98	0.84	0.84
	SALAMANDRA-2B†	41.68	0.95	0.83	0.83
	+ ACAD-TRAIN	49.69	0.97	0.84	0.84
	SALAMANDRA-7B†	45.76	0.97	0.84	0.84
	+ ACAD-TRAIN	51.51	0.96	0.84	0.84
ca→es	GPT-4.1-MINI	82.93	1.0	0.89	0.8
	GPT-4.1-NANO	78.07	1.0	0.89	0.8
	GEMINI-2.0-FLASH	84.8	1.0	0.89	0.8
	GEMINI-2.5-FLASH	83.3	1.0	0.89	0.8
	LLAMA3-8B†	77.19	1.0	0.89	0.8
	GEMMA3-27B†	82.74	1.0	0.89	0.8
	MADLAD400-7B†	44.6	0.58	0.84	0.78
	SALAMANDRA-2B†	74.12	0.96	0.88	0.79
	+ ACAD-TRAIN	84.9	1.0	0.89	0.8
	SALAMANDRA-7B†	80.48	1.0	0.89	0.8
	+ ACAD-TRAIN	85.83	1.0	0.89	0.8

Table 12: Translation results for the xx→es language pairs in ACAD-BENCH dataset. Baselines are grouped into **dedicated MMNMT models**, **medium- to small-sized open-weights models** and **large-scale proprietary general models**. Models with open weights are marked with †.

Pair	Model	d-BLEU	BP	COMET	COMET-KIWI
es→fr	GPT-4.1-MINI	43.10	0.99	0.84	0.84
	GPT-4.1-NANO	42.42	1.0	0.84	0.84
	GEMINI-2.0-FLASH	48.27	0.99	0.84	0.84
	GEMINI-2.5-FLASH	46.39	0.99	0.84	0.84
	LLAMA3-8B†	41.91	1.0	0.83	0.83
	GEMMA3-27B†	44.08	0.98	0.84	0.84
	MADLAD400-7B†	45.56	1.0	0.84	0.84
	SALAMANDRA-2B†	40.16	0.97	0.82	0.82
	+ ACAD-TRAIN	48.57	0.99	0.84	0.84
	SALAMANDRA-7B†	44.3	0.98	0.84	0.84
	+ ACAD-TRAIN	50.2	0.99	0.84	0.84
es→ca	GPT-4.1-MINI	71.4	0.99	0.9	0.82
	GPT-4.1-NANO	70.61	0.99	0.9	0.82
	GEMINI-2.0-FLASH	81.29	0.99	0.9	0.82
	GEMINI-2.5-FLASH	79.46	0.99	0.9	0.82
	LLAMA3-8B†	69.34	0.99	0.89	0.81
	GEMMA3-27B†	79.03	0.99	0.9	0.82
	MADLAD400-7B†	38.99	0.57	0.84	0.81
	SALAMANDRA-2B†	65.43	0.92	0.88	0.81
	+ ACAD-TRAIN	80.42	0.98	0.9	0.82
	SALAMANDRA-7B†	74.93	0.98	0.9	0.81
	+ ACAD-TRAIN	81.58	0.99	0.9	0.82

Table 13: Translation results for the es→xx language pairs in ACAD-BENCH dataset. Baselines are grouped into **dedicated MMNMT models**, **medium- to small-sized open-weights models** and **large-scale proprietary general models**. Models with open weights are marked with †.