Tailoring Machine Translation for Scientific Literature through Topic Filtering and Fuzzy Match Augmentation

Thomas Moerman[1], **Tom Vanallemeersch**[2], **Sara Szoc**[2] and **Arda Tezcan**[1]

[1]Language and Translation Technology Team (LT^3), Ghent University
[2]CrossLang

Correspondence: {thomas.moerman,arda.tezcan}@ugent.be, {tom.vanallemeersch,sara.szoc}@crosslang.com

Abstract

To enhance the accessibility of scientific literature in multiple languages and facilitate the exchange of information among scholars and a wider audience, there is a need for high-performing specialized machine translation (MT) engines. However, this requires efficient filtering and the use of domain-specific data. This study examines whether translation quality improves when we increase training data through combining two methods: (1) data selection via topic filtering to identify relevant sentences from larger corpora, and (2) more efficient use of data by exploiting fuzzy matches (similar translations to a given input). We apply these techniques both to sequence-to-sequence MT models and off-the-shelf multilingual large language models (LLMs) in three scientific disciplines, namely neuroscience, climatology and mobility. Our results suggest that the combination of topic filtering and FM augmentation is an effective strategy for training neural machine translation (NMT) models from scratch, not only surpassing baseline NMT models but also delivering improved translation performance compared to smaller LLMs in terms of the number of parameters. Furthermore, we find that although FM augmentation through in-context learning generally improves LLM translation performance, limited domain-specific datasets can yield results comparable to those achieved with additional multi-domain datasets.

1 Introduction

The use of a lingua franca like English for scholarly communication is, on the one hand, beneficial as it facilitates knowledge dissemination to a certain extent in the international research landscape. On the other hand, it leads to inequalities among researchers (in terms of understanding and writing) and scientific information written in different languages reaches a limited audience (Ramírez-Castañeda, 2020; Bitetti and Ferreras, 2017). Machine translation (MT) is an important support for mitigating this problem and improving knowledge dissemination. For instance, with the support of MT systems, providing translations of abstracts, keywords, and full articles could become standard practice for research programs spanning multiple languages (Amano et al., 2021). More broadly, adopting translation as a standard practice could improve access to scientific research for scientists, students, educators, policymakers, journalists, and society as a whole (Steigerwald et al., 2022).

Translating scientific texts is challenging due to specialized terminology, complex syntax, domain-specific discourse, and the fluid boundaries of scientific disciplines (Byrne, 2014). Moreover, these unique characteristics of scientific literature and the limited language resources for training MT systems further complicate the task for such systems. In the Translations and Open Science project (Fiorini et al., 2023), which we refer to henceforth as *TaOS*, custom MT engines were trained for various scientific disciplines. This effort showed that it is challenging to collect parallel training data for scientific disciplines, as many texts are only available in one language (translation is not an activity that is habitually applied in scholarly communication because of disciplinary standards and because there is a shortage of human resources).

In this paper, we approach the scarcity of scientifically oriented parallel training data for the English→French language direction by (1) applying data selection (using topic-based classifiers) to efficiently filter larger corpora in order to identify potentially relevant training material for building sequence-to-sequence neural MT (NMT) models from scratch, and (2) exploiting fuzzy match (FM) augmentation techniques (i.e. leveraging the translation of sentences similar to a given input) to make more efficient use of the available data for

Proceedings of the Eleventh Workshop on Patent and Scientific Literature Translation (PSLT 2025), pages 13–26, June 24, 2025

NMT models as well as off-the-shelf LLMs. In this study, we mainly focus on training NMT models from scratch and less to LLMs for various reasons: LLMs (the training data of which are typically unknown) may present data leakage and thus suffer from unrepresentative evaluations; they require more substantial computational infrastructure for inference (NMT models can run on CPU, whereas this is far less obvious for LLM models); and finally, the answers of instruct variants of LLMs need some post-processing. Therefore, we do not provide a comprehensive comparison between (pre-trained) NMT models and LLMs: while fine-tuning pre-trained NMT models and LLMs are common and effective, we merely focus on out-of-the-box translation capabilities of LLMs (i.e. zero-shot) and through in-context learning.

2 Related Research

2.1 NMT and LLMs in Specialized Domains

Advancements in NMT, driven mainly by adopting the transformer architecture (Vaswani et al., 2017a), have greatly enhanced translation quality across various domains. In recent years, further improvements have been achieved in MT performance with LLMs, which leverage extensive training data and advanced architectures and enhance translation accuracy, fluency, and adaptability across diverse contexts. While LLMs have consistently outperformed traditional models in general-domain MT tasks in recent years, such as for news, literary texts, and social media (Kocmi et al., 2023, 2024), their effectiveness in specialized domains remains less conclusive. In the WMT 2024 patent translation task, transformer-based NMT systems from previous years (2019 and 2020) achieved the best translation performance as measured by automatic evaluation metrics for multiple language pairs (Higashiyama, 2024). Furthermore, a recent study by Wassie et al. (2025) shows that even large fine-tuned LLMs underperform compared to transformer-based multilingual encoder-decoder models when trained on domain-specific medical translation data. These findings highlight the importance of assessing the performance of NMT systems alongside LLMs for domain-specific MT, as traditional NMT models may still offer competitive or even superior performance in such scenarios.

2.2 Science-oriented Data and NMT Models

To address the lack of parallel data for scientific texts in underrepresented European languages, the SciPar corpus was created and made publicly available through the ELRC-SHARE repository (Roussis et al., 2022). Additionally, as part of the TaOS project we mentioned earlier, Fiorini et al. (2023) compiled 316,701 parallel sentences across three scientific disciplines: (i) Climatology and Climate Change (code PE10 in the European Research Council nomenclature), (ii) Neuroscience and Disorders of the Nervous System (LS5), and (iii) Human Mobility, Environment, and Space (SH7). The sentences originated from several publication types, such as journal articles, journal article abstracts and thesis abstracts). In a more recent effort, Roussis et al. (2024) collected approximately 11 million sentence pairs for English-Spanish, English-Portuguese, and English-French from 62 academic repositories, covering Cancer Research, Energy Research, Neuroscience, Transportation Research, and general academic texts (this dataset is not publicly available).

Efforts have also been made to develop MT systems for scientific literature. In the TaOS project, NMT engines were trained for the three abovementioned scientific disciplines for the language directions English→French and French→English. The engines were trained using a combination of publicly available corpora covering a variety of domains and the abovementioned compiled sentences. The results were evaluated by various *personas*, i.e. professional translators, researchers, and students without specific knowledge of the disciplines in question. Both automatic and human evaluation showed that the specialized engines have a substantially better translation quality than the baseline, i.e. engines merely trained on the public corpora. Similarly, Roussis et al. applied domain adaptation by fine-tuning a pre-trained NMT model (OPUS-MT) for the language directions Spanish→English, Portuguese→English, and French→English, demonstrating that scientific texts enhance MT performance according to automatic evaluation metrics.

2.3 Data Selection for NMT

In order to increase the amount of domain-specific training data, various data selection approaches can be applied to corpora that cover a variety of topics. An overview of data selection techniques for do-

main adaptation in NMT can be found in Chu and Wang (2018). One possible approach is to create classifiers that are trained on data belonging to a domain (positive examples) and data not belonging to it (negative examples) and to extract potentially relevant training sentences from other corpora, as illustrated by Defauw et al. (2019). Other potential approaches consist of comparing sentences between the multi-domain corpora and domain-specific resources using metrics like embedding similarity, see e.g. Pourmostafa et al. (2021), or the application of topic clustering to sentences, for instance using Latent Dirichlet Allocation (Blei et al., 2001).

2.4 FM Augmentation for NMT and LLMs

Numerous approaches have been implemented to enhance domain-specific NMT performance by leveraging FMs from bilingual resources in the given domain. Some methods modify transformer-based architectures by adjusting the decoding process (Cao and Xiong, 2018; Reheman et al., 2023), integrating lexical memory into the NMT architecture (Feng et al., 2017), introducing additional attention layers to capture information from translation memories (TMs) (He et al., 2021), or proposing a new architecture that can effectively edit FMs to produce MT output (Bouthors et al., 2023). FMs have also been effectively integrated into NMT through data augmentation; they leverage source text similarity to retrieve FMs and incorporate them by augmenting the source sentences in training, validation and test datasets (Xu et al., 2020; Tezcan et al., 2021). This approach has proven particularly effective in specialized domains starting from training sets of approximately 300K sentence pairs, with further improvements observed for larger datasets (Tezcan et al., 2024).

FM augmentation approaches do not only enhance NMT performance. LLMs have also shown the ability to leverage FMs in domain-specific scenarios through in-context learning: highly similar FMs are added to a given input sentence in LLM prompts, enabling the LLM to replicate previously observed translation patterns (Moslem et al., 2023a; Mu et al., 2023). Furthermore, incorporating FMs (Moslem et al., 2023b) or randomly selected examples from domain-specific datasets (Alves et al., 2023) into the fine-tuning process, alongside input prompts has been shown to enhance the MT performance of LLMs.

3 Methodology

3.1 Data Selection

We performed topic filtering by applying science-oriented classifiers that extract potentially relevant training sentences from other corpora. A classifier determines how likely a target-language sentence originates from a scientific discipline. We used the FastText tool[1] to create classifiers based on the target-language part of discipline-specific TaOS training data (the positive examples consist of a random sample of sentences from one discipline, and the negative examples originate from the two other disciplines). When applying the classifier to an unseen sentence, we required a minimal score to accept it as an example of the class. This score is the lowest score observed at the best trade-off point of the ROC curve for the training examples, i.e. the point where the formula $TPR-FPR$ (true positive ratio minus false positive ratio) reaches its maximum.

We applied the classifiers to corpora covering various scientifically oriented and other topics. Given that the target-language sentences satisfy the minimal score, we retrieved the corresponding source-language sentences from the corpora and obtained additional sentence pairs to be used as NMT training data.

3.2 FM Augmentation

For FM retrieval, we followed the neural fuzzy repair (NFR) approach of Tezcan et al. (2021).[2] Given a bilingual dataset consisting of source/target sentence pairs S, T, for each source sentence $s_i \in S$ with the translations $\{t_1, \ldots, t_n\} \in T$, we retrieved the n the most similar source sentences in the same dataset $\{s_1, \ldots, s_n\} \in S$ (i.e., these are FMs), where $s_i \notin \{s_1, \ldots, s_n\}$ (i.e. we excluded exact matches), given that the FM similarity score is above a fixed threshold: $\lambda \geq 0.5$. To this end, we measured the FM score $FM(s_i, s_j)$ between two source sentences s_i and s_j as the cosine similarity between their sentence embeddings e_i and e_j:

$$FM(s_i, s_j) = \frac{e_i \cdot e_j}{\|e_i\| \times \|e_j\|} \quad (1)$$

where $\|e\|$ is the magnitude of vector e.

We generated the sentence embeddings using sent2vec (Pagliardini et al., 2018), while we effi-

[1] https://fasttext.cc/
[2] https://github.com/lt3/nfr

ciently retrieved FMs using a FAISS index (Johnson et al., 2021). The hyperparameters for sentence embedding generation and FAISS index construction are detailed in Appendices A.2 and A.3, respectively. Before FM retrieval, all sentences were segmented into subwords using SentencePiece (Kudo and Richardson, 2018), more specifically using the XLM-RoBERTa (base) tokenizer.[3] Table 1 illustrates the FM retrieval process.

S	We **found three** studies for inclusion in the review.
$score$	0.9309
FM_S	We **identified nine eligible** studies for inclusion in the review.
FM_T	Nous avons **identifié neuf** études **éligibles** pour l'inclusion dans la revue.
T	Nous avons **trouvé trois** études pour l'inclusion dans la revue.

Table 1: An example of FM retrieval for the English→French language direction in the neuroscience discipline for a given source sentence S and the reference translation T. FM_S and FM_T refer to the source and target sides of the retrieved FM with the FM similarity score, which is indicated as $score$. The non-matching parts are marked in bold.

The work of Tezcan et al. (2021) demonstrated that, in the context of transformer-based NMT systems, the augmentation of a given source sentence with the best FM yields notable improvements in MT performance but the effectiveness of incorporating additional FMs is less clear. Following this work, for the NMT systems, FM augmentation was implemented using (only) the best FM (i.e. FM with the highest similarity score), where FM-augmented source sentences S^* consist of the original source sentence, concatenated by the translation of the retrieved FM, using a separator token (S <*sep*> FM_T). The training data consists of both the original and the FM-augmented source/target sentence pairs S, T and S^*, T, respectively. During inference (i.e. on the test and validation sets), each source sentence is augmented using the same FM retrieval method described earlier. If no FMs are retrieved above the threshold $\lambda \geq 0.5$, the original (non-augmented) source sentence is used as input to the FM-augmented NMT model.

FM augmentation for LLM experiments is implemented by adding n-best FM_S/FM_T pairs to the instruction prompts to leverage the in-context learning abilities of the given LLM alongside the input sentence S for which the MT output is produced. The prompt templates are provided in Appendix A.6.

[3] https://huggingface.co/docs/transformers/v4.22.2/en/model_doc/xlm-roberta#overview

4 Experimental Setup

4.1 Data

We randomly split the TaOS data[4] for the three disciplines into training, validation, and test sets, ensuring that there is no overlap between them in terms of sentence pairs. The maximum number of tokens per sentence (prior to sub-word tokenization) in all partitions was limited to a maximum of 100. Additionally, sentences consisting of a single token were removed in the validation and test sets. Finally, we ensured that there were no unaligned sentence pairs (i.e. sentence pairs with very low translation equivalence) by analyzing the source-target pairs in the validation and test sets using the SentenceTransformer model LaBSE[5] (Feng et al., 2022) setting a minimum equivalence score of 0.6. The number of sentence pairs in the final partitions are provided in Table 2 while the average token count for each dataset can be found in Appendix A.1.

	Train	Validation	Test
Neuroscience	98,857	1,552	1,543
Climatology	95,694	1,630	1,609
Mobility	106,282	1,784	1,752

Table 2: The number of sentences partitioned from the TaOS data as training, validation and test sets per discipline.

In order to generate additional MT training data, we first created a classifier for each of the three disciplines in the TaOS data, using the method described in 3.1. We then applied these classifiers to the French sentences in the three below multi-domain corpora, filtered out low-scoring sentences, and retrieved their English pendant to obtain a set of sentence pairs:

- SciPar:[6] a collection of parallel corpora from scientific abstracts;
- EuroPat:[7] a parallel corpus of European patent data;

[4] The data supporting the findings of this study are available upon request by contacting the corresponding author(s). The data are provided for research purposes only.

[5] https://huggingface.co/sentence-transformers/LaBSE

[6] https://opus.nlpl.eu/ELRC-5067-SciPar/en&es/v1/ELRC-5067-SciPar

[7] https://europat.net/ and https://opus.nlpl.eu/EuroPat/en&fr/v3/EuroPat

- ParaCrawl:[8] a parallel data set extracted from a large set of downloaded web pages.

Before applying the classifiers, we filtered out additional sentences from the three datasets using the following approaches:

- We filtered out sentences with a low translation equivalence using the LaBSE model setting a minumum equivalence score of 0.6, as the construction of these corpora involved automated alignment, which sometimes leads to sentence pairs that are not or are only partially equivalent[9]. We only applied the equivalence detection to a 10M sample of ParaCrawl because of the high computation cost; therefore, the topic filtering is only applied to this sample.
- We filtered out short sentences (less than 10 words).

The number of sentence pairs used from the additional datasets are provided in Table 3.

	Europat	**ParaCrawl**	**SciPar**
Original	11,032,300	9,765,499	1,063,329
TF Neurosci.	2,156,482	2,508,710	392,037
TF Climat.	6,998,414	2,713,013	474,472
TF Mobility	2,610,923	5,879,689	334,144

Table 3: The number of sentence pairs used as additional training data for the NMT systems and for FM augmentation (for both NMT and LLM experiments), obtained from three datasets, before and after topic filtering (indicated as *Original* and *TF*, respectively), per discipline.

4.2 NMT Models

We trained NMT models from scratch, using configurations varying on two aspects: (i) training data and (ii) FM augmentation. All systems utilized validation sets for the given scientific discipline (i.e. neuroscience, climatology, or mobility).

Regarding the first aspect, we tested the following training data configurations:

- *1d*: TaOS data for a given discipline;
- *3d*: all TaOS data (i.e. combination of all three disciplines);
- *3d+Ext*: all TaOS data combined with all extra (i.e. *original*) multi-domain datasets (i.e. ParaCrawl, EuroPat and SciPar);
- *3d+ExtTF*: all TaOS data combined with the results of topic filtering (*TF*), as described in Section 3.1, on the extra datasets for the given discipline.

Regarding the second aspect, the above configurations were combined with FM augmentation[10], as described in Section 3.2. This increases the size of the training data for all configurations.

An overview of the training set sizes of the configurations is provided in Appendix A.4. All the NMT systems trained from these datasets utilized the transformer architecture (Vaswani et al., 2017b) and the OpenNMT-py toolkit[11] (Klein et al., 2017). Prior to training, all sentences were segmented into sub-words using SentencePiece, as described in Section 3.2. The resulting vocabulary sizes per system are provided in Appendix A.4. All systems were trained with shuffled training datasets and early stopping with 10 validation rounds in terms of accuracy and perplexity. All training runs were initialized with the same seed. For the systems that do not utilize FM augmentation, the maximum source and target lengths were set to 200 tokens. The maximum source length was doubled to 400 tokens for the systems that utilize FM augmentation. Other details regarding the hyper-parameters used for training the NMT systems are provided in Appendix A.5.

4.3 LLMs

We utilized LLMs in zero-shot and FM-augmented settings through in-context learning. We tested four models: Mistral 7B (base)[12] and 24B (instruct)[13] (Jiang et al., 2023), Tower 7B (instruct)[14] (Alves et al., 2024), which was fine-tuned on Mistral for translation-related tasks, and Mistral Nemo 12B (instruct).[15] The *instruct* variants were necessary in case of the larger models, as they proved more suitable for translation tasks without additional fine-tuning steps.

We tested two types of prompting strategies: (i) a zero-shot setting with a simple translation instruction, and (ii) a 12-shot setting, following the work of (2023a), in which prompts were augmented with

[8] https://paracrawl.eu/ and https://opus.nlpl.eu/ParaCrawl/corpus/version/ParaCrawl

[9] Based on the test sets held out from the TaOS data, it appeared that virtually all sentences minimally had this score.

[10] As an exception, due to the limited size of the *1d* training sets, we did not apply further FM augmentation for this configuration.

[11] https://github.com/OpenNMT/OpenNMT-py, v. 3.5.1.

[12] https://huggingface.co/mistralai/Mistral-7B-v0.3

[13] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501

[14] https://huggingface.co/Unbabel/TowerInstruct-Mistral-7B-v0.2

[15] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

12-best FMs (source/target pairs), as described in Section 3.2. To test the usefulness of FM augmentation in different data configurations, FMs were retrieved from the four different training datasets used for NMT training, i.e. *1d*, *3d*, *Ext* and *ExtTF*.

4.4 Evaluation

We made use of the automated evaluation metrics SacreBLEU,[16] (Post, 2018), chrF (Popović, 2015), and COMET[17] (Rei et al., 2020) to assess the quality of the (detokenized) MT output. To verify whether differences between the automated quality metric scores of the different MT systems are statistically significant, we used bootstrap resampling tests (Koehn, 2004). We performed both the automated evaluations and bootstrap resampling tests using the MATEO toolkit[18] (Vanroy et al., 2023), with the default settings for each metric.

5 Results

5.1 NMT Models

Table 4 provides the automated evaluation results for the translations generated by the different MT models on the discipline-specific test sets.

Examining the NMT models, we observe that increasing the training set size from single-discipline datasets (*1d*) to utilizing all available data from the three scientific disciplines, along with additional out-of-domain data (*3d_Ext*), positively impacted translation performance. Furthermore, applying topic filtering to the out-of-domain datasets (*ExtTF*) and incorporating FM augmentation (*FM*) further enhanced the automatic metric scores across all datasets and disciplines, highlighting the effectiveness of both techniques. The best-performing system leveraged the combined datasets from all three scientific disciplines with topic-filtered extra data (*3d+ExtTF_FM*) and FM augmentation, achieving statistically significant improvements over all other configurations. Notably, regarding the NMT experiments, FM augmentation proved most effective when paired with the dataset configuration that leveraged topic-filtered multi-domain datasets, delivering greater improvements than when applied to the full, unfiltered datasets.

[16] https://github.com/mjpost/sacrebleu v. 2.4.1. (SacreBLEU and chrF)
[17] https://huggingface.co/Unbabel/wmt22-comet-da
[18] https://mateo.ivdnt.org/

5.2 LLMs

When analyzing the results across different LLMs, we can make multiple observations. Firstly, FM augmentation through in-context learning enhanced the performance of all tested LLMs, across all metrics and disciplines, with one exception: Mistral Nemo 13B model generally achieved the highest automatic metric scores in zero-shot setting without FM augmentation. Secondly, the impact of the additional datasets is inconclusive for the Tower, Nemo, and Mistral 24B models, while FM augmentation improved performance. Expanding the pool of sentences in the limited discipline-specific datasets (i.e. *1d*) for FM retrieval, whether by merging training data from all three disciplines or incorporating additional multi-domain datasets with or without topic filtering, did not consistently lead to improvements or, at best, resulted in only marginal gains. For instance, the Mistral 24B model achieved the highest COMET scores in the climatology and mobility disciplines utilizing only the discipline-specific datasets for FM augmentation. It could be argued that given the additional computational resources required for extracting FMs from more extensive data sets, restricting the pool of sentences for FM retrieval to the given scientific discipline (using approximately 100K sentences) offers a more favourable balance between efficiency and quality. Please also see Figure 1 for an overview of the best-performing configuration per model and discipline.

When comparing different LLMs, we observe that the translation performances of the general-purpose models (e.g., Mistral, Mistral Nemo) improved with the increasing number of parameters. The Tower 7B models deviated from this general pattern, outperforming the smaller Mistral 7B models using the same data configurations, as well as the larger Nemo 13B models across all metrics and disciplines. These results confirm the effectiveness of the Tower model compared to other LLMs of similar size in the MT task (Kocmi et al., 2024). In the context of LLM parameter size, it should also be highlighted that although the larger models generally resulted in higher scores, the smallest Mistral model (7B) achieved the highest relative gains from FM augmentation compared to the zero-shot setting. For instance, in the neuroscience discipline, *Mistral 7B_FM_3d+ExtTF* outperforms *Mistral 7B* by +2.78 COMET, whereas *Mistral*

NMT model	Neuroscience			Climatology			Mobility		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
1d	39.11	65.15	79.28	29.57	57.99	76.05	30.14	59.45	76.98
3d	43.11	68.40	83.06	35.24	62.62	81.01	33.96	62.32	81.73
3d_FM	43.67	68.74	83.47	35.38	62.55	81.14	33.96	62.40	81.73
3d+Ext	44.40	69.42	84.70	35.70	63.35	82.48	36.11	64.01	84.88
3d+Ext_FM	44.78	69.58	85.12	36.32	63.54	82.79	36.09	63.96	84.80
3d+ExtTF	44.99	69.75	84.73	36.28	63.76	82.54	36.89	64.49	84.96
3d+ExtTF_FM	**46.33**‡	**70.58**‡	**85.30**†	**36.97**†	**64.00**†	**82.82**	**37.68**‡	**64.81***	**85.27**‡
LLM									
Mistral 7B	32.85	61.58	81.64	28.66	57.98	79.97	29.98	59.19	82.48
Mistral 7B_FM_1d	39.74	65.94	84.37	32.76	60.71	82.45	33.55	61.87	85.01
Mistral 7B_FM_3d	39.35	65.71	84.29	32.69	60.56	82.35	33.70	61.88	85.04
Mistral 7B_FM_3d+Ext	**40.72**	66.23	84.37	**34.79***	**61.90***	**82.76**	**35.35**	**62.73**	**85.11**
Mistral 7B_FM_3d+ExtTF	40.50	**66.28**	**84.42**	34.42	61.52	82.70	35.27	62.69	85.08
Tower 7B	40.81	66.55	84.74	34.28	61.92	82.77	36.18	63.19	85.02
Tower 7B_FM_1d	41.97	67.11	84.92	35.22	62.28	82.92	35.84	63.02	85.30
Tower 7B_FM_3d	41.98	67.11	84.90	35.03	62.16	82.84	35.92	63.08	85.36
Tower 7B_FM_3d+Ext	**43.17***	**67.74***	**84.95**	**36.68***	**62.91**	82.90	**37.08**	63.55	85.22
Tower 7B_FM_3d+ExtTF	42.82	67.53	84.93	36.43	62.84	**82.96**	36.99	**63.62**	**85.38**
Nemo 13B	40.04	**67.01***	**84.77**†	**33.58**	**62.31**†	**82.97**†	**34.55**	**63.16**†	**85.44**†
Nemo 13B_FM_1d	40.63	65.84	83.95	33.16	60.54	81.95	33.78	61.29	84.46
Nemo 13B_FM_3d	40.72	65.98	84.04	33.27	60.77	82.15	33.47	61.07	84.30
Nemo 13B_FM_3d+Ext	40.94	66.08	83.95	33.39	60.44	81.64	34.16	61.21	84.05
Nemo 13B_FM_3d+ExtTF	**41.05**	66.16	83.93	33.42	60.57	81.72	33.72	60.99	84.01
Mistral 24B	42.23	68.49	85.51	35.90	63.70	83.61	37.46	65.07	86.18
Mistral 24B_FM_1d	44.88	69.81	86.10	37.27	64.44	**84.13**	38.72	65.68	**86.72***
Mistral 24B_FM_3d	44.94	69.90	**86.18**	37.24	64.43	84.12	38.57	65.54	86.64
Mistral 24B_FM_3d+Ext	**45.08**	69.91	86.11	37.26	64.47	84.12	**38.85**	65.68	86.61
Mistral 24B_FM_3d+ExtTF	45.05	**69.92**	**86.18**	**37.31**	**64.51**	84.12	**38.85**	**65.69**	86.62

Table 4: Results of the automatic evaluations performed for the different MT systems, per discipline. For each model (i.e. per section), the highest metric scores are highlighted in bold and statistically significant improvements are denoted by $*$, †, and ‡, representing $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively, based on the lowest p values obtained when compared to all other configurations of the same model type.

24B_FM_ExtTF shows an improvement of +0.67 COMET over *Mistral 24B*.

5.3 Cross-comparison

In a final analysis, we compare the performance of the best configuration per model type using COMET as the primary evaluation metric, with Figure 1 presenting the automated evaluation results per discipline of the single best-performing setup for each model type. This figure further includes the statistical significance of the performance differences observed between the various model types.

Upon reviewing the overall best-performing model, we observe that the largest LLM, Mistral 24B, surpassed all other models with respect to COMET scores, achieving an improvement of up to +1.26 within the mobility discipline compared to all other tested models. However, in the neuroscience discipline, the highest BLEU and chrF scores were attained by the top-performing NMT configuration (*3d+ExtTM_FM*), with improvements of +1.28 BLEU and +0.66 chrF compared to the best-performing LLM (Mistral 24B). Moreover, the best NMT configuration surpassed Mistral 7B, Tower 7B and Nemo 13B across all disciplines and metrics, with the exception of COMET scores in the climatology and mobility disciplines, where Tower 7B achieved higher scores. Since BLEU and chrF emphasize token and character overlap between the MT output and the reference translations, it can be hypothesized that while the NMT model is better at maintaining discipline-specific lexical choices for the neuroscience domain, the COMET scores, which measure semantic similarity, suggest that Tower 7B and Mistral 24B better capture the overall meaning. However, this hypothesis requires validation through manual evaluation and error analysis of MT performance in subsequent studies.

6 Conclusions and Future Work

Developing highly accurate MT systems for specialized scientific disciplines continues to be a significant challenge due to unique textual characteristics and the scarcity of language resources neces-

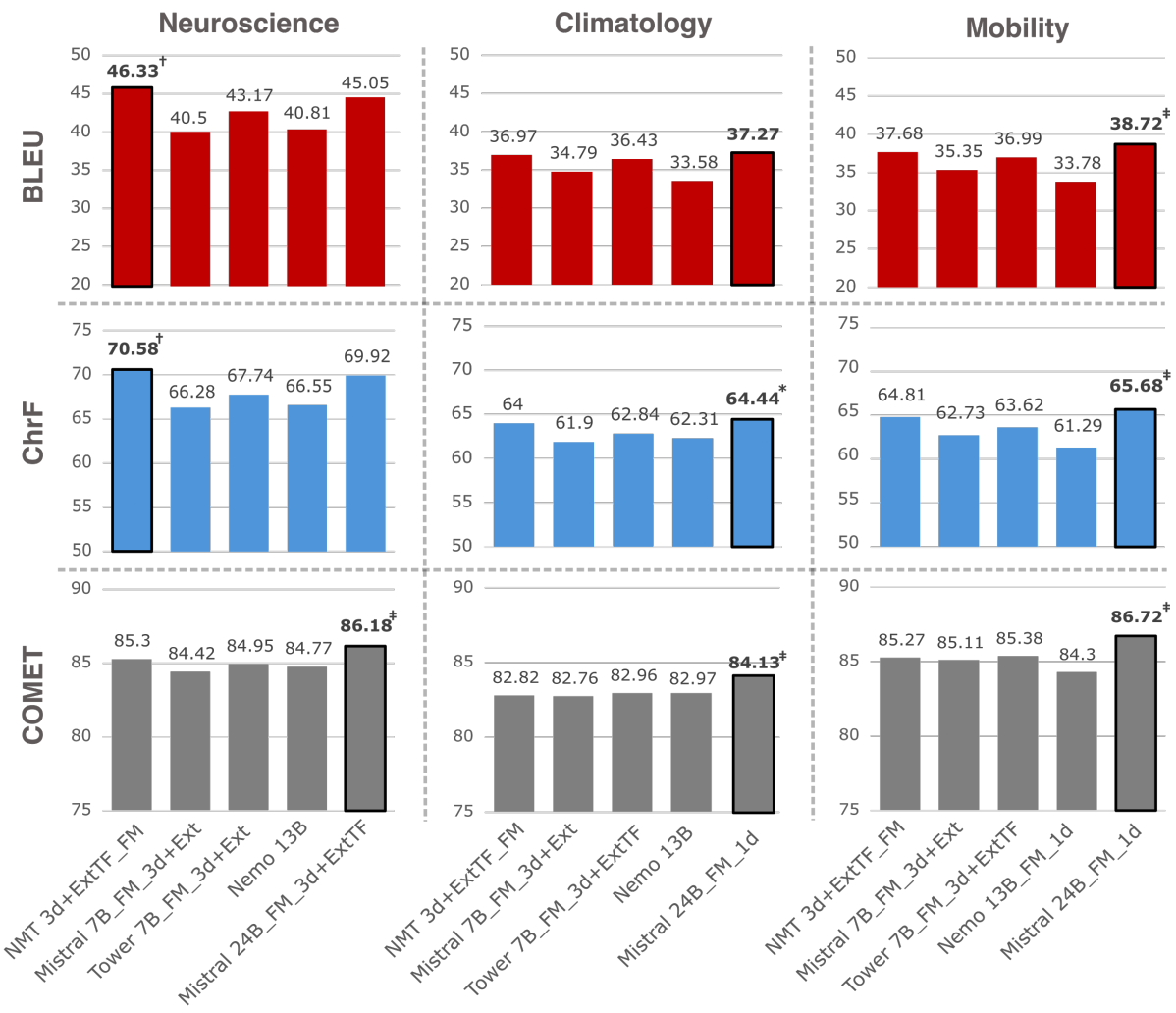

Figure 1: Results of the automatic evaluations for the best-performing configuration per model type selected in terms of COMET scores (NMT vs. LLMs), per discipline. The highest metric scores per metric and discipline are highlighted in bold and statistically significant improvements are denoted by *, †, and ‡, representing $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively, based on the lowest p values obtained when compared to all other models.

sary for building effective MT systems.

In this study, we combined two existing methodologies, aiming to tailor MT systems for the scientific domain, namely topic filtering of large, multi-domain datasets to extract relevant NMT training data and FM augmentation to utilize the available data more efficiently. To this end, we trained NMT models from scratch and employed four LLMs to evaluate their zero-shot and in-context learning capabilities. Our experiments, which covered three scientific disciplines, namely neuroscience, climatology, and mobility, in the English→French language direction, revealed that combining topic filtering with FM augmentation effectively enhances NMT models trained from scratch. However, although FM augmentation via in-context learning proved beneficial for most of the LLMs tested, the value of additional datasets in this context, regardless of whether they included topic filtering, remained inconclusive. Our findings suggest that smaller, discipline-specific datasets could yield comparable results to larger datasets when employed for FM augmentation in this specific setting, while incurring significantly lower computational costs.

Furthermore, our findings enable a comparison between NMT models trained from scratch and LLMs (without further fine-tuning) for this task. We demonstrated that specialized NMT models can achieve superior translation performance compared to out-of-the-box LLMs in this discipline-specific scenario, with these improvements being more pronounced when compared to smaller LLMs with fewer parameters. Therefore, it can be argued that

these improvements in translation quality and other benefits, such as reduced inference costs, make NMT systems a viable option for translating scientific literature, particularly when computational resources are limited. However, given the positive correlation we observed between translation performance and the increasing number of LLM parameters, our findings suggest that larger LLMs, even in the absence of further fine-tuning, could deliver better translation performance than such specialized NMT models.

In future studies, we will test additional configurations with given datasets, for example, retrieving FMs for the test/validation sets only from the discipline-specific datasets while using extra, larger datasets as additional NMT training data and for FM augmentation on the training set. Moreover, we will investigate the effectiveness of additionally fine-tuning pre-trained NMT models and LLMs using the in-domain datasets, with or without FM augmentation (i.e. zero- vs. few-shot settings), as both approaches have been shown to further improve MT performance in previous studies.

7 Limitations

One of the main limitations of this study is its limited scope in terms of MT experiments, which do not explore fine-tuning strategies of pre-trained NMT models or LLMs. Moreover, our experiments were limited to automatic assessment of MT performance, which may not fully reflect translation quality, and to a single language pair, albeit across three scientific disciplines. Human evaluation of MT performance and additional experiments in different language directions would be necessary to validate our findings. Furthermore, our evaluation focused on the effectiveness of combining specific data selection and augmentation methods rather than comparing them against a wider range of alternative approaches. Finally, we did not explore the efficiency of different n values for integrating n-best FMs into the LLM prompts or additional prompting strategies.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), which is funded by Ghent University, FWO and the Flemish Government department EWI.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.17733.

Tatsuya Amano, Violeta Berdejo Espinola, Alec P. Christie, Kate Willott, Munemitsu Akasaka, András Báldi, Anna Berthinussen, Sandro Bertolino, Andrew J. Bladon, Min Chen, Chang-Yong Choi, Magda Bou Dagher Kharrat, Luis G. de Oliveira, Perla Farhat, Marina Golivets, Nataly Hidalgo Aranzamendi, Kerstin Jantke, Joanna Kajzer-Bonk, M. Çisel Kemahlı Aytekin, Igor Khorozyan, Kensuke Kito, Ko Konno, Da-Li Lin, Nick Littlewood, Yang Liu, Yifan Liu, Matthias-Claudio Loretto, Valentina Marconi, Philip Martin, William H. Morgan, Juan P. Narváez-Gómez, Pablo Jose Negret, Elham Nourani, Jose M. Ochoa Quintero, Nancy Ockendon, Rachel Rui Ying Oh, Silviu Petrovan, Ana C. Piovezan-Borges, Ingrid L. Pollet, Danielle L. Ramos, Ana L. Reboredo Segovia, A. Nayelli Rivera-Villanueva, Ricardo Rocha, Marie-Morgane Rouyer, Katherine A. Sainsbury, Richard Schuster, Dominik Schwab, Çağan H. Şekercioğlu, Hemin Seo, Gorm Shackelford, Yushin Shinoda, Rebecca K. Smith, Shan-dar Tao, Ming-shan Tsai, Elizabeth Tyler, Flóra Vajna, José Osvaldo Valdebenito, Svetlana Vozykova, Paweł Waryszak, Veronica Zamora-Gutierrez, Rafael D. Zenni, Wenjun Zhou, and William J. Sutherland. 2021. Tapping into non-english-language science for the conservation of global biodiversity. *bioRxiv*.

Mario Santiago Di Bitetti and Julián A. Ferreras. 2017. Publish (in english) or perish: The effect on citation rate of using languages other than english in scientific publications. *Ambio*, 46:121–127.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Maxime Bouthors, Josep Crego, and François Yvon. 2023. Towards example-based NMT with multi-Levenshtein transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1830–1846, Singapore. Association for Computational Linguistics.

Jody Byrne. 2014. *Scientific and technical translation explained: A nuts and bolts guide for beginners*. Routledge.

Qian Cao and Deyi Xiong. 2018. Encoding gated translation memory into neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3042–3047.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Arne Defauw, Tom Vanallemeersch, Sara Szoc, Frederic Everaert, Koen Van Winckel, Kim Scholte, Joris Brabers, and Joachim Van den Bogaert. 2019. Collecting domain specific data for MT: an evaluation of the ParaCrawl pipeline. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 186–195, Dublin, Ireland. European Association for Machine Translation.

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.

Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark. Association for Computational Linguistics.

Susanna Fiorini, Arda Tezcan, Tom Vanallemeersch, Sara Szoc, Kristin Migdisi, Laurens Meeus, and Lieve Macken. 2023. Translations and open science: exploring how translation technologies can support multilingualism in scholarly communication. In *Proceedings of the International Conference HiT-IT 2023*, pages 41–51. INCOMA Ltd.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online. Association for Computational Linguistics.

Shohei Higashiyama. 2024. Results of the WAT/WMT 2024 shared task on patent translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 118–123, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

J. Johnson, M. Douze, and H. Jegou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(03):535–547.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *Computing Research Repository*, arXiv:1701.02810.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere,

Finland. European Association for Machine Translation.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. Fine-tuning large language models for adaptive machine translation. *ArXiv*, abs/2312.12740.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana, 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Valeria Ramírez-Castañeda. 2020. Disadvantages in preparing and publishing scientific papers caused by the dominance of the english language in science: The case of colombian researchers in biological sciences. *PLoS ONE*, 15.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. 2022. Scipar: A collection of parallel corpora from scientific abstracts. In *International Conference on Language Resources and Evaluation*.

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).

Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck. 2021. Selecting parallel in-domain sentences for neural machine translation using monolingual texts. *ArXiv*, abs/2112.06096.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora Y C Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future. *BioScience*, 72(10):988–998.

Arda Tezcan, Bram Bulté, and Bram Vanroy. 2021. Towards a better integration of fuzzy matches in neural machine translation through data augmentation. *Informatics*, 8(1).

Arda Tezcan, Alina Skidanova, and Thomas Moerman. 2024. Improving fuzzy match augmented neural machine translation in specialised domains through synthetic data. *Prague Bull. Math. Linguistics*, 122:9–42.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500. European Association for Machine Translation (EAMT).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Aman Kassahun Wassie, Mahdi Molaei, and Yasmin Moslem. 2025. Domain-specific translation with open-source large language models: Resource-oriented analysis. *Preprint*, arXiv:2412.05862.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Dataset	Avg. No. Tokens (Std. Dev.) English	French
Neuroscience (train)	23.8 (11.6)	27.5 (13.2)
Neuroscience (val.)	23.7 (11.7)	27.4 (13.3)
Neuroscience (test)	23.3 (11.7)	27.3 (13.7)
Climatology (train)	25.8 (12.1)	29.2 (13.7)
Climatology (val.)	26.1 (12.3)	29.5 (13.6)
Climatology (test)	26.3 (12.8)	29.4 (14.3)
Mobility (train)	26.2 (13.1)	28.3 (13.8)
Mobility (val.)	26.5 (13.2)	28.7 (14.2)
Mobility (test)	26.3 (12.8)	28.4 (13.8)
Scipar	25.1 (13.2)	27.9 (14.4)
EuroPat	30.7 (18.6)	30.9 (18.4)
ParaCrawl	22.3 (11.7)	24.4 (12.9)

Table 5: Average number of tokens per dataset prior to sub-word tokenization, with the standard deviation shown in parentheses.

A Appendix

A.1 Dataset Statistics

A.2 Sent2vec Hyper-parameters

To train sent2vec models, we used the hyper-parameters that are suggested in the description paper (Pagliardini et al., 2018) for a sent2vec model trained on Wikipedia data containing both unigrams and bigrams. The hyper-parameters values are provided in Table 6.

Hyper-parameter	Value
embedding dimension	480
minimum word count	8
minimum target word count	20
initial learning rate	0.2
epochs	9
sub-sampling hyper-parameter	5×10^{-6}
bigrams dropped per sentence	4
number of negatives sampled	10

Table 6: Hyper-parameters for training sent2vec models.

A.3 FAISS Configuration

For efficient retrieval of FMs, we created a flat FAISS index with an inverted file system (IVF) of 4096 clusters. We used cosine similarity as the match metric by adding the L2-normalized vectors of the sentence representation to the index and using an L2-normalized sentence vector as an input query. For more information on FAISS, please see https://github.com/facebookresearch/faiss/wiki.

A.4 NMT Training Data and Vocabulary Sizes

System	Neuroscience	Climatology	Mobility
1d	98,857	95,694	106,282
3d	300,833	300,833	300,833
3d_FM	601,666	601,666	601,666
3d+Ext	22,161,961	22,161,961	22,161,961
3d+Ext_FM	44,323,799	44,323,799	44,323,799
3d+ExtTF	5,358,062	10,486,732	9,125,589
3d+ExtTF_FM	10,708,321	20,969,963	18,249,664

Table 7: The total number of bilingual sentence pairs used for training the NMT systems, per discipline.

System	Lang.	Neurosci.	Climat.	Mobility
1d	src	22,216	25,049	27,101
	tgt	21,414	24,448	25,814
3d	src	32,791	32,791	32,791
	tgt	32,274	32,274	32,274
3d_FM	src	35,936	35,936	35,936
	trg	32,274	32,274	32,274
3d+Ext	src	67,995	67,995	67,995
	tgt	62,333	62,333	62,333
3d+Ext_FM	src	69,031	69,031	69,031
	tgt	62,333	62,333	62,333
3d+ExtTF	src	55,516	57,280	63,512
	tgt	51,869	53,669	58,643
3d+ExtTF_FM	src	56,506	58,314	64,248
	tgt	51,869	53,669	58,643

Table 8: Vocabulary sizes (source/target) of the NMT systems, per discipline.

A.5 NMT Hyper-parameters

Hyper-parameter	Value
source/target embedding dimension	512
size of hidden layers	512
feed-forward layers	2048
number of heads	8
number of layers	6
batch size	32
gradient accumulation	4
dropout	0.1
warm-up steps	8000
optimizer	Adam

Table 9: Common hyper-parameter values used for training the NMT systems.

We performed evaluations on a given validation set after every 10% of the training data was processed during each NMT training (i.e. 10 evaluations per epoch).

A.6 LLM Prompts

The zero-shot and in-context learning (i.e. few-shot) experiments employed different prompt templates depending on the model type. Table 10 presents the prompt templates used for the Mistral-7B-v0.3 base model, following Moslem et

al. (2023a), and for all the instruct models, following Format 1 described in Alves et al. (2023).

Model	Translation Type	Prompt Template
Base	Zero-shot	English: ⟨source_segment⟩ French:
Base	Few-shot (e.g., 2-shot)	English: ⟨source_fuzzy_match$_2$⟩ French: ⟨target_fuzzy_match$_2$⟩ English: ⟨source_fuzzy_match$_1$⟩ French: ⟨target_fuzzy_match$_1$⟩ English: ⟨source_segment⟩ French:
Instruct	Zero-shot	Translate the source text from X to Y. Source: ⟨source_segment⟩ Target:
Instruct	Few-shot (e.g., 2-shot)	Translate the source text from X to Y. Source: ⟨source_fuzzy_match$_2$⟩ Target: ⟨target_fuzzy_match$_2$⟩ Translate the source text from X to Y. Source: ⟨source_fuzzy_match$_1$⟩ Target: ⟨target_fuzzy_match$_1$⟩ Translate the source text from X to Y. Source: ⟨source_segment⟩ Target:

Table 10: Prompt templates used for zero-shot and few-shot translation with the different LLMs tested in this study. In the few-shot templates, fuzzy matches are ordered from the nth-most similar match to the most similar (where n refers to the number of shots), followed by the source segment to be translated.