Model Cards for the MaTOS Project

Ziqian Peng, Rachel Bawden, François Yvon

Model Cards for the MaTOS Project

Ziqian Peng, Rachel Bawden and François Yvon

August 2025

MaTOS — Livrable D3-1.1

Machine Translation for Open Science - ANR-22-CE23-0033

Model Cards for the MaTOS Project

August 2025

Table des matières

1 Introduction

This report provides a full documentation for machine translation (MT) systems trained in MaTOS [1] project for the purpose of investigating and improving the translation quality of scientific text. We applied the state-of-the-art Transformer-based neural architectures, including the vanilla Transformer [Vaswani et al., 2017] (Section 3) and mBART50 [Liu et al., 2020b, Tang et al., 2021] (Section 4). We also consider large language models (LLMs) such as TowerBase [Alves et al., 2024] derived from LLaMA2 [Touvron et al., 2023] for translation-related tasks, and EuroLLM [Martins et al., 2025]. Up to the present time, our models are trained on scientific abstracts in divers domains, including published open-source parallel corpora such as **SciPar** [Roussis et al., 2022], and datasets curated and constructed in this project such as **TAL** [2] for natural language processing (NLP) and **STEP** for Earth, environmental and planetary sciences. In addition, we examined the challenges of translating lengthy sequences using architectures of different positional encodings (PEs), such as mBART50 with learned absolute PEs (APEs), NLLB [Costa-jussà et al., 2024] with sinusoidal APEs [Vaswani et al., 2017], and TowerBase with RoPE [Su et al., 2024]. This experiment utilized parallel documents from **STEP**, and parallel pseudo-documents constructed using the TED talks corpus [Cettolo et al., 2012]. The configurations of the resulting MT systems are presented in Section 5.

We focused on the English-French (EN–FR) language pair. Section 2 gives a brief and concise presentation of corpora used to train the systems and test sets applied to perform a primary evaluation in BLEU [Papineni et al., 2002] scores. To the end, we conclude this work and discuss our future exploration in Section 7.

2 Data

2.1 Parallel corpora for pretraining

SciPar [Roussis et al., 2022] is a multilingual collection of parallel abstracts from openly published bachelor theses, master theses and doctoral dissertations across various fields. We focused on the EN-FR language pair, which consists of 1.1M parallel sentences, from which we randomly selected 3000 parallel sentences for each of the validation set and the test set.

2.2 Domain-specific parallel corpora : TAL

2.2.1 Parallel corpora for training and validation

TAL is made up of abstracts of articles and thesis in NLP, including 1701 thesis abstracts retrieved from theses.fr and 1357 article abstracts extracted from ISTEX. Given these raw documents, we first segmented them with Trankit [Nguyen et al., 2021], which better distinguishes the different features of the point in the text. Sentences were then

1. https://anr-matos.github.io/
2. https://github.com/ANR-MaTOS/Resources/tree/main/NLP-abstracts

aligned using hunalign [3] [Varga et al., 2005] to build a corpus of sentences. For holistic document-level machine translation, the parallel sentences within each document are concatenated and translated as a block. [4] We note **TAL-D** the document-level version of the **TAL** corpus and **TAL-S** the sentence-level one. No specific tag is employed to mark sentence boundaries if this is not specified.

TAL-MR : We also constructed a training set **TAL-MR** using the documents of the corpus **TAL** to apply the multi-resolution (MR) training strategy [Sun et al., 2022]. This MR approach consists in constituting a training set of balanced document length distribution, by cutting each document (from **TAL-D** in our case) into K sub-documents several times, with $K \in \{1, 2, 4, 8, \dots\}$. In other words, a document of length 8 is divided into 15 sub-parts, with one document of 8 sentences, 2 sub-documents of 4 sentences, 4 sub-documents of 2 sentences and 8 sub-documents of one sentence. The augmented set contains 35376 segment pairs instead of 2858, and the average sentence length is reduced to 74 / 87 for EN/FR instead of 236 / 276 tokens.

2.2.2 Test sets

THE (dev and test) contains two collections of 101 and 100 abstracts in the field of TAL randomly extracted from theses.fr without overlap with **TAL**. **rTAL** contains 246 parallel abstracts of articles published in the TAL journal.

These articles are aligned at sentence level using the same method as for **TAL** ; they have also been filtered using TransQuest [Ranasinghe et al., 2020] and the alignments have been manually reviewed. The statistics in table 1 describe these different corpora.

Another test sets **IWSLT** consisting of 53 pseudo-documents from 10 transcriptions of oral presentations for IWSLT 2023 [Salesky et al., 2023] are also constructed with balanced length distribution, to evaluate the impact of document length. Please refer to [Peng et al., 2024b] for more details about this experiment.

	SciPar			**TAL**				
	train.	valid.	test	train.	valid.	**THE**	**rTAL**	**IWSLT**
Nb. sentences	1116325	3000	3000	2858	101	100	246	53
avg. length in a doc.	37	38	37	265	317	327	129	402
avg. length of sentences in a doc.	-	-	-	34	35	33	32	24

TABLE 1 – Statistics of dataset in the field of NLP, including the training set (train.), validation est (valid.) and test sets. The length are computed with respect to sentence pieces of raw texts in source language using the BPE model of MBART50 (1-M).

3. https://github.com/danielvarga/hunalign

4. The raw data from theses.fr are collected by Maxime Bouthors in 2022, that of ISTEX are collected by Mathilde Huguin in the MaTOS project.

2.3 Domain-specific parallel corpora : STEP

This section introduces our parallel corpora in the field of STEP : the Earth, environmental and planetary sciences ("Sciences de la Terre, de l'Environnement et des Planètes" in French). Our datasets are constructed using parallel scientific abstracts in the EN-FR language pair, collected from :

— THE : theses.fr
— CRAS : Comptes Rendus de l'Académie des Sciences - Earth and Planetary Science
— CRG : Comptes Rendus Géoscience
— CanMin : Canadian Mineralogist
— CJES : Canadian Journal of Earth Sciences
— ISTEX : ISTEX
— BSGF : (Bulletins de la Société Géologique de France)

All documents are segmented using Trankit, aligned using BertAlign [Liu and Zhu, 2022] and filtered with respected to alignment scores evaluated by TransQuest. Table 2 summarises the statistics of parallel corpora in STEP.

	STEP		**STEP-v2**		test sets			
	train.	valid.	train.	valid.	**THE**	**CRAS**	**BSGF**	**CRG**
#doc	8382	300	10063	400	100	100	132	59
avg. #sent in a doc.	8	8	8	8	13	7	10	6
avg. length in a doc.	352	373	341	348	518	283	474	261
avg. length of sentences in a doc.	47	45	46	43	41	43	48	43

TABLE 2 – Statistics of dataset in the field of STEP, including the training set (train.), validation est (valid.) and test sets. The length are computed with respect to sentence pieces of raw texts in source language using the BPE model of MBART50 (1-M).

2.3.1 Parallel corpora for training and validation

STEP is a set of parallel documents for training and validation, Table 3 presents its statistics and distribution of document resources.

	all	THE	CRAS	CanMin	CJES	ISTEX
train.	8382	1217	1846	798	4521	-
valid.	300	100	100	-	100	-

TABLE 3 – Statistics of **STEP**

STEP-v2 is an updated version of **STEP**. It includes 1781 additional parallel abstracts from ISTEX. Table 4 reports the updated statistics and distribution of document resources.

	all	THE	CRAS	CanMin	CJES	ISTEX
train.	10,063	1217	1846	798	4521	1681
valid.	400	100	100	-	100	100

TABLE 4 – Statistics of **STEP-v2**

Similar to the corpus **TAL**, we denote **STEP-D** and **STEP-D-v2** the document-level version of **STEP** and **STEP-v2** respectively.

2.3.2 Test sets

We received four test sets constructed in the MaTOS project for the STEP field, please refer to Table 5 for detailed information on their resources and statistics.

	#doc	mean	min	max	#sent
THE	100	518	198	1010	1295
CRAS	100	283	66	678	676
CRG	59	261	107	562	368
BSGF	132	473	64	1354	1308

TABLE 5 – Statistics of test sets in STEP for the EN-FR language pair, with the amount of parallel abstracts (#doc), the amount of corresponding parallel sentences (#sent), the average, minimum and maximum source document length in pieces (mean, min and max). The length are computed with respect to sentence pieces of raw texts using the BPE model of MBART50 (1-M).

2.4 Parallel corpora for length issues : TED

To systematically investigate the challenges related to input lengths in translating long documents [Bao et al., 2021, Li et al., 2023, Lupo et al., 2023], we have constructed a series of parallel pseudo-documents using TED talks [Cettolo et al., 2012]. We summarise in the subsections 2.4.1 and 2.4.2 the data preparation procedure, and report the data statistics in Table 6, with more details presented in the Appendix of [Peng et al., 2025].

2.4.1 Parallel corpora for training and validation

Our training set consists of pseudo-documents from both the training and validation splits of **IWSLT-2016**. [5] Our goal is to simulate real corpora of parallel documents with source documents shorter than a certain length l_{max} – using $l_{max} = 1024$. We split all document pairs whose source side is longer than 1024 tokens into fragments. [6]

5. https://wit3.fbk.eu/2016-01

6. All the statistics in tokens in Section 2.4 is counted using the tokeniser of NLLB [Costa-jussà et al., 2024]. All the full TED talks in our corpora start with the title, then the description and the talk

	TED-full		TED-G		TED-U	
	train	dev	train	dev	train	dev
Count	1831	19	15625	160	10582	106
Length	2915	2861	341	339	504	512

	sent	256	512	768	1024	1200	1600	2048	doc
Count	5103	503	261	184	142	123	100	80	52
Length	23	233	450	638	827	955	1175	1468	2259

TABLE 6 – Left : Statistics of the TED talks training and dev sets. Right : Statistics of the TED talks test sets from IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**. 'Count' denotes the number of parallel pseudo-documents, 'Length' denotes the average length of source (i.e. English) pseudo-documents (in NLLB tokens).

TED-G For each document pair, we iterate the following procedure : (1) sample a maximum pseudo-document length l'_i following the same Gaussian-like length distribution as the full TED talks with $l'_i < l_{max}$; (2) concatenate consecutive sentence pairs up to l'_i to form a training pseudo-documents s_i.[7] The development set is built similarly, using document pairs from IWSLT **tst2010** and **tst2011**. We denote these training datasets as **TED-G** (G for Gaussian).

TED-U We also consider another dataset generation strategy, to produce a more balanced length distribution, for which we do as above but we sample uniformly : $l'_i \sim U(128, l_{max})$. Fine-tuning with the resulting **TED-U** corpus allows us to contrast two corpora with differences in document length distributions.

2.4.2 Test sets

To evaluate MT systems for their ability to handle documents of varying context window sizes and extrapolate beyond the training sample lengths, we build a series of test sets of increasing document lengths. For each document in IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**, we accumulate consecutive sentence pairs into parallel pseudo-documents such that all resulting source texts have a length close to l_{max}, with $l_{max} \in \{256, 512, 1024, 1200, 1600, 2048\}$.[8] Contrarily to training sets, test sets are homogeneous in length. Evaluation is always performed with complete original talks, after concatenating and aligning all the corresponding parts.

3 Translation Systems for Scientific Abstracts (Transformer)

Our experiments on vanilla Transformer are implemented with the `fairseq` [Ott et al., 2019] framework. All models are based on the Transformer_base architecture, with 6

before being split into pseudo-documents. Tags `<description>` and `<title>` are removed.

7. If concatenating the last sentence pair (x_n, y_n) into the current pseudo-document pair exceeds l_{max}, (x_n, y_n) will yield a single parallel sequence, to respect the maximum length l_{max} in our training datasets.

8. At the end of each talk, we concatenate the last parallel sentences into the last pseudo-document if they are shorter than 50 to avoid exceedingly short parallel sequences.

layers, 8 attention heads, hidden size of 512 and feed forward size of 2048. The maximum position of each input sequence is limited to 4096. The max token size of each batch is 4096 for the baseline Baseline-TR and 2048 for others. We updated the parameters every four batches for Baseline-TR and two for others. Our systems are all trained on NVIDIA RTX A5000, with a patience value of 5 to stop training if the most recent 5 epochs cannot improve the BLEU score on the validation set.

model name	backbone	dataset	bsz	freq.	max pos. (src/tgt)	misc.
Baseline-TR	-	**SciPar**	4096	4	4096/4096	Sent2Sent
FT-TR TAL-S	Baseline-TR	**TAL-S**	2048	2	4096/4096	Sent2Sent
FT-TR TAL-D	Baseline-TR	**TAL-D***	2048	2	4096/4096	Doc2Doc
FT-TR TAL-D-MR	Baseline-TR	**TAL-D-MR***	2048	2	4096/4096	Doc2Doc
FT-TR TAL-D Mask-Past	Baseline-TR	**TAL-D***	2048	2	4096/4096	Doc2Doc
FT-TR TAL-D Mask-Future	Baseline-TR	**TAL-D***	2048	2	4096/4096	Doc2Doc
FT-TR TAL-D Mask-All	Baseline-TR	**TAL-D***	2048	2	4096/4096	Doc2Doc

Table 7 – Experiment configurations for models based on vanilla transformer (the Transformer_base architecture in `fairseq`), including the pre-trained model (backbone) for fine-tuning, the parallel corpus for fine-tuning (dataset), the batch size (bsz) in tokens, the update frequency (freq.) by batch, the limit of max position (max pos.) set to filter source and target sequences, and other notes (misc.). * indicates that we use `<sep>` to mark sentence boundaries.

The experiment configurations are reported in table 7, where FT-TR TAL-D Mask-Past, FT-TR TAL-D Mask-Future and FT-TR TAL-D Mask-All denotes translation systems fine-tuned from Baseline-TR, with the past, future, or all source contexts sentences masked respectively. In the case of masking future context, we force the decoder generating at least the same amount of `<sep>` as the input source document before the end of generation. Please refer to [Peng, 2023] for more details about the evaluation and analysis of these models.

4 Translation Systems for Scientific Abstracts (mBART50)

In this section, we present our translation engines based on mBART50 (1-M), which is a pretrained multilingual translation system of encoder-decoder architecture derived from BART through continue training with multilingual parallel corpora from English to 49 other languages [Liu et al., 2020b, Tang et al., 2021]. We followed a two-stage fine-tuning strategy. More precisely, we first fine-tuned mBart50 (1-M) with parallel sentences to develop a sentence-level translation system (FT-SciPar), then fine-tuned this model for document-level translation.

All experiments presented in this section are implemented with the `fairseq` [Ott et al., 2019] framework. All models are based on the `mbart_large` architecture, with 12 layers for each of the encoder and decoder, 16 attention heads, hidden size of 1024 and feed forward size of 4096. They are trained using a GPU `NVIDIA RTX A6000 48G` and 12 CPU with 8G memory each. The position embedding of BART models is lear-

ned during training. To avoid overfitting, we use an early stopping procedure with the value of `patience` equal to 5 (epochs), depending on the BLEU scores evaluated on the validation set. For decoding with `fairseq-interactive`, we use the parameter values `max-len-a`=1.5 instead of the default value (1.2), and `max-len-b`=10. These two parameters are used to control the length of the sentences generated by translation. The beam size is set to 5.

Please check table 8 for a list of translation systems based on MBART50 (1-M) and their specific configurations.

model name	backbone	dataset	bsz	freq.	max pos. (src/tgt)	misc.
FT-SciPar	MBART50(1-M)	**SciPar**	4096	4	1024/1024	Sent2Sent
FT-TAL-S	FT-SciPar	**TAL-S**	2048	2	1024/1024	Sent2Sent
FT-TAL-D	FT-SciPar	**TAL-D**	2048	2	1024/1024	Doc2Doc
FT-TAL-MR	FT-SciPar	**TAL-MR**	2048	2	1024/1024	Doc2Doc
FT-STEP-D (V1)	FT-SciPar	**STEP-D**	4096	2	2048/2048	Doc2Doc
FT-STEP-D-ISTEX (V2)	FT-SciPar	**STEP-D-v2**	4096	2	2048/2048	Doc2Doc

TABLE 8 – Experiment configurations for models based on MBART50(1-M), including the pre-trained model (backbone) for fine-tuning, the parallel corpus for fine-tuning (dataset), the batch size (bsz) in tokens, the update frequency (freq.) by batch, and the max position (max pos.) for source and target sequences, and other notes (misc.).

Results and analysis of models trained on **TAL** are published in [Peng et al., 2024b]. Regarding models specified in **STEP**, we report ds-BLEU [9] [Peng et al., 2024b] in table 9 to give a first insight of their performance. [10]

4.1 Baseline : FT-SciPar

To begin, we fine-tuned MBART50 (1-M) on **SciPar**, a collection of cross-domain parallel sentences extracted from scientific abstracts (Section 2.1). During training, the batch size is set to 4096, and we updated the parameters every four batches. The maximum position of each input sequence is limited to 1024. We consider the resulting model FT-SciPar as our sentence-level baseline. Other in domain translation systems in this section are fine-tuned based on it.

9. We evaluate ds-BLEU using SacreBLEU [Post, 2018] with signature : nrefs :1|case :mixed|eff :yes|tok :13a|smooth :exp|version :2.4.0.

10. In table 9, we reported the translation performance of TowerBase-7B and TowerInstruct-7B-v0.1 to compare with our custom models. For TowerBase, we applied the zero-shot prompt « English : SRC\n French : » given only the source input `SRC`. Regarding TowerInstruct, we applied the chat template with the following instruction : « Translate the following text from English into French.\n English : SRC\n French : » also in zero-shot. We have evaluated their performance using three shots and five shots on the validation set of **TAL-D**, while the zero-shot performance of both models is better than few shot in our experiments.

	THE	CRAS	CRG	BSGF
FT-STEP-D	46.4 (0.98)	34.9 (0.98)	47.7 (0.98)	40.3 (0.98)
FT-STEP-D-v2	46.4 (0.98)	34.6 (0.97)	49.0 (0.98)	41.7 (0.98)
DeepLPro	49.7 (0.99)	34.4 (0.98)	49.3 (0.98)	46.3 (0.99)
SystranPro	46.1 (0.98)	32.1 (0.98)	45.3 (0.98)	43.1 (0.98)
TowerBase	45.5 (0.98)	33.1 (0.97)	48.0 (0.97)	38.1 (0.96)
TowerInstruct	44.7 (0.99)	31.7 (0.98)	44.5 (0.98)	40.5 (0.98)

Table 9 – Scores of ds-BLEU (BP) of FT STEP-D and FT STEP-D-v2 evaluated on test sets in the field of STEP, compared with the pro version of the commercial systems DeepL and Systran, along with TowerBase-7B and TowerInstruct-7B-v0.1 [Alves et al., 2024], which are translation systems derived from LLaMA2.

4.2 Sentence-level systems

4.2.1 FT-TAL-S

For comparison with document-level models, we trained FT-TAL-S, a sentence-level translation system derived from FT-SciPar via supervised fine-tuning on **TAL-S** datasets presented in Section 2.2.1. The batch size is set to 2048, and we updated the parameters every two batches. The maximum position of each input sentence is 1024.

4.3 Document-level systems

4.3.1 FT-TAL-D

We denote FT-TAL-D the MT model fine-tuned from FT-SciPar on parallel documents (abstracts in NLP field in this case) of **TAL-D**. FT-TAL-D is able to take an entire abstract as input and give its translation in an holistic way. Except the setting in common presented at the beginning of Section 4, to train FT-TAL-D, we have set the batch size as 2048 and the parameters were updated every two batches. The maximal position is also 1024.

4.3.2 FT-TAL-D MR

FT-TAL-D is the document-level MT system obtained by fined-tuning FT-SciPar on **TAL-MR** to explore the effectiveness of multi-resolution training [Sun et al., 2022] that is shortly presented in Section 2.2. Except the training set, all the configuration is the same as FT-TAL-D.

4.4 Document-level systems with position extension

The **STEP** training set comprises several document pairs longer than 1024, potentially due to the domain shift between the training data of the BPE model for MBART50 (1-M) and abstracts in **STEP**, which contains mathematics formula, specific terminologies, etc. Under this context, we extend the maximum position from 1024 to 2048 when fine-tuning on **STEP**. In other words, we initialise the position embedding with dimension 2048, and copy the pretrained parameter of dimension 1024 to the first 1024 position reserved for input sequences. The following subsections describe two stable versions of document-level MT systems trained in this scenario.

4.4.1 FT-STEP-D (V1)

FT-STEP-D (version 1) is a document-level MT model that fine tune FT-SciPar on parallel abstracts from the corpus **STEP** in the domain of Earth, environmental and planetary sciences. As presented in Section 2.3, this training set contains more than 8k parallel documents. In this case, we set the batch size as 4096 and we updated the parameters every two batches.

4.4.2 FT-STEP-D-ISTEX (V2)

Given the updated parallel abstracts **STEP-v2** (cf. Section 2.3), we have trained a second version of document-level translation system in the field of STEP, and we denoted this model as FT-STEP-D-ISTEX (or FT-STEP-D-v2). The training protocol and configurations is exactly the same as FT-STEP-D.

5 Investigating Length Issues through Positional Encodings

We further observed challenges related to length, particularly position bias towards the final part of input sequences, when applying Transformer-based models to document-level MT tasks [Peng et al., 2024b]. One reason of these issues is the unbalanced positional encodings in the training corpus, that favours smaller position indices while underfitts larger ones [Peng et al., 2024a, Zhu et al., 2024, Peng et al., 2025]. Therefore, we explored how to mitigate this problem using more balanced positional encoding distributions, through a) a better positional encoding strategy for APEs that maps the PE distribution of a training corpus to a uniform distribution, denoted as UNIFPE (see [Peng et al., 2025, Section 4]), or b) a training corpus of a more balanced document length distribution (i.e. **TED-U**). We also proposed a systematic evaluation method to measure the ability of MT systems to process documents of increasing lengths. This exploration is presented in Peng et al. [2025].

This section provides the experiment configurations for all MT models resulting from this experiment, with Table 10 giving an overview of these configurations.

It concerns three different architectures, including MBART50 with learned APEs [Tang et al., 2021], NLLB with sinusoidal APEs [Costa-jussà et al., 2024] and TOWER-BASE [Alves et al., 2024] with RoPE [Su et al., 2024]. Since MBART50 relies on learned

APEs, integrating unifPE to uniformise the APE distribution of all the batches resulted in severe translation quality degradation due to catastrophic forgetting. One simple but effective solution is to incorporate unifPE only in a portion of batches, as reported in Table 10. In contrast, our preliminary experiments show that integrating unifPE to all batches instead of a percentage of batches while fine-tuning TowerBase or NLLB converges and results in better performance. Therefore, we fixed the ratio of unifPE as 1 if we apply it to these models. In our experiments, we set up the target maximum length that controls the unifPE algorithm to 2048 for mBART50 and NLLB models and to 4096 for the decoder-only TowerBase models, and we performed experiments on the corpora **STEP-D-v2**, **TED-G** and **TED-U** (see Sections 2.3 and 2.4).

model name	backbone	PE	unifPE	dataset	bsz	freq.	max pos.	archi.
FT STEP-D-v2	FT-SciPar	Learned APE	0	**STEP-D-v2**	4096*	2	2048/2048	enc-dec
Unif-BART-STEP-r5	FT-SciPar	Learned APE	0.5	**STEP-D-v2**	4096*	2	2048/4096	enc-dec
Unif-BART-STEP-r7	FT-SciPar	Learned APE	0.7	**STEP-D-v2**	4096*	2	2048/4096	enc-dec
FT-Tower-STEP	TowerBase	RoPE	0	**STEP-D-v2**	8	4	4096	dec-only
Unif-Tower-STEP	TowerBase	RoPE	1	**STEP-D-v2**	8	4	4096	dec-only
FT-NLLB-G	NLLB	Sinusoidal APE	0	**TED-G**	32	4	2048/2048	enc-dec
Unif-NLLB-G	NLLB	Sinusoidal APE	1	**TED-G**	32	4	2048/2048	enc-dec
FT-NLLB-U	NLLB	Sinusoidal APE	0	**TED-U**	32	4	2048/2048	enc-dec
Unif-NLLB-U	NLLB	Sinusoidal APE	1	**TED-U**	32	4	2048/2048	enc-dec
FT-Tower-G	TowerBase	RoPE	0	**TED-U**	8	2	4096	dec-only
Unif-Tower-G	TowerBase	RoPE	1	**TED-G**	8	2	4096	dec-only
FT-Tower-U	TowerBase	RoPE	0	**TED-U**	8	2	4096	dec-only
Unif-Tower-U	TowerBase	RoPE	1	**TED-U**	8	2	4096	dec-only

Table 10 – Experiment configurations for models presented in Section 5. These configurations include the pre-trained backbone model (backbone), the positional encoding (PE), the ratio of unifPE for fine-tuning (unifPE), the parallel corpus for fine-tuning (dataset), the batch size (bsz) counted in tokens (with *) or in number of input sequences, the update frequency (freq.) by batch, and the maximum position (max pos.) for source and target sequences (source/target), and the model architecture (archi.). NLLB denotes nllb200-distilled-600M. TowerBase denotes TowerBase-7B.

5.1 unifPE STEP (mBART50)

We first applied unifPE to fine tune the mBART50-based model **FT SciPar** on **STEP-D-v2**, aiming to improve MT engines introduced in Section 4.4.

5.1.1 Baseline : FT-STEP-D-v2

We consider the FT-STEP-D-v2 in Section 4.4.2 as our baseline model, which is developed through document-level fine-tuning.

5.1.2 Unif-mBART-STEP-r5

The MT system UNIF-MBART-STEP-r5 fine-tuned **FT SciPar** using the same configurations as FT-STEP-D-V2, but incorporating the UNIFPE algorithm to 50% of the fine-tuning batches.

5.1.3 Unif-mBART-STEP-r7

Similarly, the model UNIF-MBART-STEP-r7 applied UNIFPE algorithm to 70% of the fine-tuning batches. All the other experiment settings are the same as that of FT-STEP-D-V2.

5.2 unifPE STEP (Tower)

Subsequently, we examined the effect of UNIFPE on an LLM-based architecture, TOWERBASE-7B [11] [Alves et al., 2024] (TOWERBASE for short), derived from Llama2 using translation-related tasks. TOWERBASE uses RoPE [Su et al., 2024] to encode relative PEs, which nonetheless encodes some form of APE signal in some dimensions [Peng et al., 2024a], and may therefore be also mildly impacted by the PE training distribution.

5.2.1 Baseline : FT-Tower-STEP

We performed supervised fine-tuning using QLoRA [Dettmers et al., 2023] and bfloat16 on TOWERBASE as our baseline model, denoted as FT-TOWER-STEP. The prompt for fine-tuning is « Translate the following text from English into French.\nEnglish : SRC\nFrench : TGT ». Since TOWERBASE relies on a decoder-only architecture, the maximum position doubles to 4096 in our experiments to account for both the source and the target sequence, no matter what is the effective maximum position observed from the training corpus. The batch size is 8 with 4 gradient accumulation steps. The learning rate is $2e-4$ adjusted by a *cosine* schedule, without warm-up steps nor packing. We fine-tuned the model for one epoch. The inference is performed without additional in-context examples, with bfloat16 and greedy search.

5.2.2 Unif-Tower-STEP

We applied UNIFPE to all batches during the fine-tuning process using QLoRA on the **STEP-D-v2** corpus, following the same configurations as FT-TOWER-STEP. We refer to the resulting MT system as UNIF-TOWER-STEP.

5.3 unifPE TED-G & TED-U (NLLB)

We considered NLLB200-DISTILLED-600M [12] or NLLB for short [Costa-jussà et al., 2024] as a representative encoder-decoder model based on APEs. NLLB is a 12-layer

11. https://huggingface.co/Unbabel/TowerBase-7B-v0.1
12. https://huggingface.co/facebook/nllb-200-distilled-600M

	THE	**CRAS**	**CRG**	**BSGF**
FT STEP-D-v2	46.4 (0.98)	34.6 (0.97)	49.0 (0.98)	41.7 (0.98)
UNIF-MBART-STEP-r5	46.7 (0.98)	35.0 (0.97)	48.6 (0.98)	41.8 (0.98)
UNIF-MBART-STEP-r7	46.6 (0.98)	35.2 (0.98)	49.0 (0.98)	40.5 (0.98)
TOWERBASE	45.5 (0.98)	33.1 (0.97)	48.0 (0.97)	38.1 (0.96)
FT-TOWER-STEP	47.1 (0.98)	35.0 (0.98)	48.9 (0.98)	39.7 (0.98)
UNIF-TOWER-STEP	46.8 (0.98)	35.1 (0.98)	49.0 (0.98)	39.7 (0.98)

TABLE 11 – ds-BLEU scores (and BP) of MT systems fine-tuned on the **STEP-D-v2** corpus, evaluated on test sets in STEP and compared with TOWERBASE.

model name	200	300	400	500	600	700
FT STEP-D-v2	42.8 (0.98)	39.4 (0.98)	34.4 (0.97)	27.5 (0.95)	17.6 (0.94)	4.2 (0.96).
UNIF-BART-STEP-r5	43.2 (0.97)	40.1 (0.95)	35.8 (0.93)	28.1 (0.89)	19.3 (0.85)	6.4 (0.81)
UNIF-BART-STEP-r7	43.2 (0.96)	40.3 (0.94)	36.3 (0.92)	29.1 (0.87)	20.7 (0.83)	8.3 (0.81)
TOWERBASE	41.3 (0.94)	41.0 (0.95)	36.6 (0.95)	35.5 (0.88)	33.0 (0.89)	31.5 (0.81)
FT-TOWER-STEP	45.8 (0.98)	45.4 (0.98)	44.6 (0.98)	43.6 (0.98)	38.7 (0.98)	37.6 (0.99)
UNIF-TOWER-STEP	45.9 (0.98)	45.5 (0.98)	45.0 (0.98)	44.2 (0.98)	42.6 (0.98)	41.2 (0.98)

TABLE 12 – ds-BLEU (and BP) evaluated on documents d translated at final positions (i.e. $d'd$) following a document d' of increasing length $l \in \{200, 300, 400, 500, 600, 700\}$ [Peng et al., 2024b, Section 4.3]. The documents d is the set of **THE** in **STEP**. We take 5 different d' from **BSGF** that satisfies the length constraint and this table reports their average scores.

model name	200	300	400	500	600	700
FT STEP-D-v2	45.9 (0.98)	45.8 (0.99)	45.9 (0.99)	45.9 (0.98)	45.6 (0.99)	45.8 (0.99)
UNIF-BART-STEP-r5	46.1 (0.98)	46.0 (0.98)	46.0 (0.98)	45.9 (0.98)	45.6 (0.98)	45.8 (0.98)
UNIF-BART-STEP-r7	46.1 (0.98)	45.9 (0.98)	45.7 (0.98)	45.6 (0.98)	45.2 (0.97)	45.4 (0.97)
TOWERBASE	40.7 (0.89)	38.1 (0.85)	35.4 (0.80)	33.1 (0.78)	37.1 (0.86)	26.7 (0.65)
FT-TOWER-STEP	46.7 (0.98)	46.1 (0.97)	45.9 (0.98)	45.6 (0.98)	45.6 (0.98)	45.0 (0.98)
UNIF-TOWER-STEP	46.4 (0.98)	45.9 (0.97)	45.7 (0.98)	45.6 (0.98)	45.2 (0.98)	45.0 (0.98)

TABLE 13 – ds-BLEU (and BP) evaluated on documents d translated at initial positions (i.e. dd') followed by a document d' of increasing length $l \in \{200, 300, 400, 500, 600, 700\}$ [Peng et al., 2024b, Section 4.3]. The documents d is the set of **THE** in **STEP**. We take 5 different d' from **BSGF** that satisfies the length constraint and this table reports their average scores.

encoder-decoder multilingual MT model pre-trained in 200 languages. We used the *HuggingFace* implementation, which relies on sinusoidal APEs [Vaswani et al., 2017]. Please refer to [Peng et al., 2025] for the evaluation results.

5.3.1 Baseline : FT-NLLB-G

FT-NLLB-G is the baseline model for our experiments performed on NLLB. To build it, we fine-tuned the pretrained model on **TED-G** with learning rate $5e-4$, 500 warm-up steps, 4 parallel pseudo-documents per batch and 32 gradient accumulation steps. An early stopping criterion with a patience of 5 epochs is also applied, according to the d-BLEU [Liu et al., 2020a] evaluated on the validation set. For inference on test sets, the beam size is set to 5 and the batch size is set to 4.

5.3.2 Unif-NLLB-G

Unif-NLLB-G is the MT model fined-tuned from NLLB following the same configurations as FT-NLLB-G, but integrated the unifPE algorithm to all batches during fine-tuning.

5.3.3 FT-NLLB-U

To explore the impact of document length distributions, we fine-tuned NLLB using the same experiment settings as FT-NLLB-G but on the **TED-U** corpus, and we named the resulting model as FT-NLLB-U.

5.3.4 Unif-NLLB-U

To examine the impact of unifPE on a corpus of a more balanced document length distribution, we fine-tuned NLLB exactly as Unif-NLLB-G but on the **TED-U** corpus, to obtain the MT engine Unif-NLLB-U.

5.4 unifPE TED-G & TED-U (Tower)

This section outlines the experiment settings used to fine-tune TowerBase. For results analysis and performance comparisons, please refer to [Peng et al., 2025].

5.4.1 Baseline : FT-Tower-G

Similar to Section 5.2, we performed supervised fine-tuning using QLoRA [Dettmers et al., 2023] and bfloat16 on **TED-G** to build the baseline model, with the same prompt pattern and the same maximum position as FT-Tower-STEP. The resulting MT system is referred to as FT-Tower-G. The batch size is 8 with 2 gradient accumulation steps. The learning rate is $2e-5$ adjusted by a *cosine* schedule, without warm-up steps nor packing. We fine-tuned the model for two epochs and saved checkpoints every 50 steps in the second epoch. We then chose the checkpoint with the best d-BLEU [Liu et al., 2020a] on the validation set. The inference is performed without additional in-context examples, with bfloat16 and greedy search.

5.4.2 Unif-Tower-G

The system UNIF-TOWER-G is derived from TOWERBASE following the same fine-tuning strategy as FT-TOWER-G in Section 5.4.1, while it applied UNIFPE during fine-tuning to uniformise the absolute PEs.

5.4.3 FT-Tower-U

FT-TOWER-U is the MT system obtained through the same fine-tuning process as that of FT-TOWER-G, but using the **TED-U** corpus.

5.4.4 Unif-Tower-U

In addition, we have fine-tuned TOWERBASE exactly as UNIF-TOWER-G, but on the **TED-U** corpus to obtain the MT system UNIF-TOWER-U.

5.5 Preliminary evaluation results

This subsection outlines the performance of MT engines introduced in Section 5. For models fine-tuned on **STEP-D-v2**, as reported in Table 11, models fine-tuned with our without UNIFPE show comparable performance in ds-BLEU on the four test sets in the STEP field. These models are further evaluated using the $d'd$ and dd' methods introduced in Peng et al. [2024b, Section 4.3] for their capacity to translate the initial part (Table 13) and final part of input sequences (Table 12). We report also the scores of TOWERBASE for comparison.[13] Results show that the fine-tuned MT models are robust to noisy information in the dd' scenario, while the translation quality of the final parts of lengthy sequences degrades when the sequence length gets longer. The UNIFPE algorithm helps to mitigate the problem for both MBART50 and TOWERBASE.

However, for models fine-tuned on **TED-G** and **TED-U**, we only observed such improvement brought by UNIFPE for NLLB models but not for TOWERBASE, according to results in Peng et al. [2025]. We assume that one potential reason is that parallel abstracts in STEP fields represent a more semantically complexe texts than pseudo-documents constructed using TED talks for document-level MT tasks.

6 Translation Systems for Scientific Abstracts (LLMs)

We have also fine-tuned recent LLMs, including TOWERBASE [Alves et al., 2024] (see Section 5) and EuroLLM-9B (EUROLLM)[14] [Martins et al., 2024, 2025], on our parallel abstracts **TAL-D** and **STEP-D-v2**. This section presents the experiment settings for the fine-tuned models, with Table 14 presents their performance on our test sets of parallel abstracts, evaluated in ds-BLEU.

13. For TOWERBASE, we applied the zero-shot prompt « English : SRC\n French : » given only the source input `SRC`.

14. https://huggingface.co/utter-project/EuroLLM-9B

6.1 FT-Tower-TAL

We fine-tuned TowerBase on **TAL-D** for two epochs, using QLoRA [Dettmers et al., 2023] and bfloat16. The batch size and gradient accumulation steps are 8 and 2 respectively. We apply a *cosine* learning rate schedule, with learning rate set to $2e-4$, without warm-up steps nor packing. The prompt for fine-tuning is « Translate the following text from English into French.\nEnglish : SRC\nFrench : TGT ». The inference use the same prompt with bfloat16 and greedy search. The resulting model is referred to as FT-Tower-TAL.

6.2 FT-Euro-TAL

We fine-tuned EuroLLM on **TAL-D** following the same configurations as FT-Tower-TAL, and we refer to the fine-tuned model as FT-Euro-TAL.

6.3 FT-Tower-STEP

please see the Section 5.2.1 for more details about FT-Tower-STEP.

6.4 FT-Euro-STEP

We fine-tuned EuroLLM on **STEP-D-v2** with almost the same hyper-parameter settings as FT-Tower-TAL. The differences are 1) the model is fine-tuned for one epoch, and 2) we set the gradient accumulation steps as 4 instead of 2. The resulting MT system is denoted as FT-Euro-STEP.

	STEP			
	THE	**CRAS**	**CRG**	**BSGF**
SystranPro	46.1 (0.98)	32.1 (0.98)	45.3 (0.98)	43.1 (0.98)
DeepLPro	49.7 (0.99)	34.4 (0.98)	49.3 (0.98)	46.3 (0.99)
TowerBase	45.5 (0.98)	33.1 (0.97)	48.0 (0.97)	38.1 (0.96)
FT-Tower-STEP	47.1 (0.98)	35.0 (0.98)	48.9 (0.98)	39.7 (0.98)
EuroLLM	49.3 (0.98)	35.5 (0.97)	51.5 (0.98)	41.8 (0.98)
FT-Euro-STEP	49.7 (0.98)	35.2 (0.97)	50.1 (0.97)	42.6 (0.98)

	TAL	
	THE	**rTAL**
SystranPro	41.6 (0.98)	31.9 (0.95)
DeepLPro	45.0 (0.98)	35.1 (0.96)
TowerBase	40.1 (0.97)	31.1 (0.95)
FT-Tower-TAL	42.1 (0.98)	33.2 (0.95)
EuroLLM	42.7 (0.98)	33.5 (0.96)
FT-Euro-TAL	43.9 (0.98)	34.0 (0.95)

Table 14 – Scores of ds-BLEU (BP) of fine-tuned LLMs. Left : FT-Tower-STEP and FT-Euro-STEP evaluated on the test sets of STEP. Right : FT-Tower-TAL and FT-Euro-TAL evaluated on the test sets of TAL. We compare their performance with the pro version of the commercial systems DeepL and Systran, along with TowerBase-7B and EuroLLM-9B.

7 Conclusion

In conclusion, this report presents the stable versions of translation systems developed within the MaTOS project, designed for two objectives : a) translating English scientific

abstracts into French and b) examining the impact of positional encoding distributions in traditional encoder-decoder architectures and decoder-only LLMs for long document translation. The models in (a) are based on the vanilla Transformer architecture, or the pretrained MBART50(1-M) model. Our baseline models learned general knowledge of scientific abstracts from the **SciPar** corpus, while our fine-tuned systems specialize in the domains of NLP or STEP. For objective (b), the translation engines enhanced translation quality by manipulating the training distribution of input lengths or position encodings. However, the challenge of translating extremely long sequences persists. Our ongoing work focuses on exploring advanced techniques, such as supervised attention and retrieval-augmented generation, to improve the translation quality of lengthy scholarly documents while maintaining the coherence and consistency in the translated texts.

Bibliographie

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. Tower : An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=EHPns3hVkj. [Cited on pages 4, 11, 12, 14, and 17.]

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. G-transformer for document-level machine translation. 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 3442–3455, Online, August 2021. Association for Computational Linguistics. doi : 10.18653/v1/2021.acl-long.267. URL https://aclanthology.org/2021.acl-long.267. [Cited on page 7.]

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3 : Web inventory of transcribed and translated talks. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy, May 28–30 2012. European Association for Machine Translation. URL https://aclanthology.org/2012.eamt-1.60. [Cited on pages 4 and 7.]

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, Jeff Wang, and NLLB Team. Scaling neural machine translation to 200 languages. *Nature*, 630(8018) :841–846, June 2024.

doi : 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x. ISBN : 1476-4687. [Cited on pages 4, 7, 12, and 14.]

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA : Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU. [Cited on pages 14, 16, and 18.]

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. P-transformer : Towards better document-to-document neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31 :38593870, September 2023. ISSN 2329-9290. doi : 10.1109/TASLP.2023.3313445. URL https://doi.org/10.1109/TASLP.2023.3313445. [Cited on page 7.]

Lei Liu and Min Zhu. Bertalign : Improved word embedding-based sentence alignment for ChineseEnglish parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2) :621–634, 12 2022. ISSN 2055-7671. doi : 10.1093/llc/fqac089. URL https://doi.org/10.1093/llc/fqac089. [Cited on page 6.]

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8 :726–742, 11 2020a. ISSN 2307-387X. doi : 10.1162/tacl_a_00343. URL https://doi.org/10.1162/tacl_a_00343. [Cited on page 16.]

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8 :726–742, 2020b. doi : 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47. [Cited on pages 4 and 9.]

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.insights-1.4. [Cited on page 7.]

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm : Multilingual language models for europe. Preprint arXiv :2409.16235, 2024. URL https://arxiv.org/abs/2409.16235. [Cited on page 17.]

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm-9b : Technical report. Preprint

arXiv :2506.04079, 2025. URL https://arxiv.org/abs/2506.04079. [Cited on pages 4 and 17.]

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. Trankit : A light-weight transformer-based toolkit for multilingual natural language processing. In Dimitra Gkatzia and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, pages 80–90, Online, April 2021. Association for Computational Linguistics. doi : 10.18653/v1/2021.eacl-demos.10. URL https://aclanthology.org/2021.eacl-demos.10. [Cited on page 4.]

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq : A fast, extensible toolkit for sequence modeling. In Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi : 10.18653/v1/N19-4009. URL https://aclanthology.org/N19-4009. [Cited on pages 8 and 9.]

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, July 2002. Association for Computational Linguistics. doi : 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040. [Cited on page 4.]

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN : Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=wHBfxhZu1u. [Cited on pages 12 and 14.]

Ziqian Peng. Document-level Machine Translation For Scientific Texts. Technical report, ISIR, Université Pierre et Marie Curie UMR CNRS 7222, September 2023. URL https://hal.science/hal-04258660. [Cited on page 9.]

Ziqian Peng, Rachel Bawden, and François Yvon. À propos des difficultés de traduire automatiquement de longs documents. In Mathieu Balaguer, Nihed Bendahman, Lydia-Mai Ho-dac, Julie Mauclair, Jose G Moreno, and Julien Pinquier, editors, *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 2–21, Toulouse, France, 7 2024b. ATALA and AFPC. URL https://aclanthology.org/2024.jeptalnrecital-taln.1. [Cited on pages 5, 10, 12, 15, and 17.]

Ziqian Peng, Rachel Bawden, and François Yvon. Investigating length issues in document-level machine translation. In Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of*

Machine Translation Summit XX : Volume 1, pages 4–23, Geneva, Switzerland, June 2025. European Association for Machine Translation. ISBN 978-2-9701897-0-1. URL https://aclanthology.org/2025.mtsummit-1.3/. [Cited on pages 7, 12, 15, 16, and 17.]

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319. [Cited on page 10.]

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. TransQuest : Translation quality estimation with cross-lingual transformers. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi : 10.18653/v1/2020.coling-main.445. URL https://aclanthology.org/2020.coling-main.445. [Cited on page 5.]

Dimitrios Roussis, Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis, and Vassilis Katsouros. SciPar : A collection of parallel corpora from scientific abstracts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2652–2657, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.284. [Cited on page 4.]

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi : 10.18653/v1/2023.iwslt-1.2. URL https://aclanthology.org/2023.iwslt-1.2. [Cited on page 5.]

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer : Enhanced transformer with rotary position embedding. *Neurocomputing*, 568 : 127063, 2024. ISSN 0925-2312. doi : https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864. [Cited on pages 4, 12, and 14.]

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics : ACL 2022*, pages 3537–3548, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi : 10.18653/v1/2022.findings-acl.279. URL https://aclanthology.org/2022.findings-acl.279. [Cited on pages 5 and 11.]

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In

Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pages 3450–3466, Online, August 2021. Association for Computational Linguistics. doi : 10.18653/v1/2021.findings-acl.304. URL https://aclanthology.org/2021.findings-acl.304. [Cited on pages 4, 9, and 12.]

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2 : Open foundation and fine-tuned chat models, 2023. [Cited on page 4.]

Dániel Varga, Péter Halaácsy, András Kornai, Voktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596, 2005. [Cited on page 5.]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf. [Cited on pages 4 and 15.]

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. PoSE : Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3Z1gxuAQrA. [Cited on page 12.]