Findings of the WMT 2024 Biomedical Translation Shared Task: Test Sets on Abstract Level

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Du Nunzio, Federica Vezzani, et al.

Findings of the WMT 2024 Biomedical Translation Shared Task: Test Sets on Abstract Level

Mariana Neves[1] * **Cristian Grozea**[2] **Philippe Thomas**[3] **Roland Roller**[3]
Rachel Bawden[4] **Aurélie Névéol**[5] **Steffen Castle**[3]
Vanessa Bonato[6] **Giorgio Maria Di Nunzio**[6] **Federica Vezzani**[6]
Maika Vicente Navarro[7] **Lana Yeganova**[8] **Antonio Jimeno Yepes**[9,10]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]Fraunhofer Institute FOKUS, Berlin, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[4]Inria, Paris, France
[5]Université Paris-Saclay, CNRS, LISN, Orsay, France
[6]Dept. of Linguistic and Literary Studies University of Padua, Italy
[7]Leica Biosystems, Australia
[8]NCBI/NLM/NIH, Bethesda, USA
[9]RMIT University, Australia
[10]Unstructured Technologies, USA

Abstract

We present the results of the ninth edition of the Biomedical Translation Task at WMT'24. We released test sets for six language pairs, namely, French, German, Italian, Portuguese, Russian, and Spanish, from and into English. Each test set consists of 50 abstracts from PubMed. Differently from previous years, we did not split abstracts into sentences. We received submissions from five teams, and for almost all language directions. We used a baseline/comparison system based on Llama 3.1 and share the source code at https://github.com/cgrozea/wmt24biomed-ref.

1 Introduction

In this paper, we present a description and the findings of the ninth edition of the Biomedical Translation Task,[1] which took place at the ninth edition of the Conference for Machine Translation (WMT'24). The shared task aims to support advances in Machine Translation (MT) in the biomedical domain, especially for scientific literature. Previous editions of the shared task addressed up to seven language pairs and included the release of training and test sets (Bojar et al., 2016; Jimeno Yepes et al., 2017; Neves et al., 2018; Bawden et al., 2019, 2020; Yeganova et al., 2021; Neves et al., 2022, 2023). All previous data is available in the shared task repository.[2]

Similar to previous years, our test sets consist of biomedical abstracts, which have been included to PubMed[3] just before publishing the test set, to decrease the likelihood of data contamination. We prepared test sets for six languages from and into English, namely, French (fr2en, en2fr), German (de2en, en2de), Italian (it2en, en2it), Portuguese (pt2en, en2pt), Russian (ru2en, en2ru), and Spanish (es2en, en2es). The test sets consist of 50 abstract pairs for each of the 12 language directions above. Some of the test sets were also released as test suites in the General Task of WMT (Kocmi et al., 2024). After the release of the test sets, the participants had around two weeks to submit their automatic translations. For this year's shared task, the following features were introduced:

- The selection of the articles for the test sets was based on topics of interest to the task organizers (Section 2);

* The contributions of the authors are the following: MN prepared the MEDLINE test sets, performed manual validation, and organized the shared task; CG developed the baseline system; PT, RR, RB, AN, SC, VB, GMN, FV, MVN, LY performed manual validation; AJY performed manual validation and the automatic evaluation, as well as co-organized the shared task; All authors approved the final version of the manuscript. E-mail for contact: mariana.lara-neves@bfr.bund.de

[1] http://www2.statmt.org/wmt24/biomedical-translation-task.html

[2] https://github.com/biomedical-translation-corpora/corpora

[3] https://pubmed.ncbi.nlm.nih.gov/

- The test sets consist of paragraphs comprising the papers' title and the abstract, i.e. no sentence splitting and alignment were carried out (Section 2);
- Consequently, we only performed a manual evaluation on the abstract level (cf. Section 6);
- We used as a baseline/comparison a local large language model, Llama 3.1 (cf. Section 3);
- We performed the automatic evaluation also based on COMET (Rei et al., 2020), besides BLEU (Papineni et al., 2002).

2 Test sets

We downloaded the daily update files from PubMed[4] around mid-April for the preparation of the test sets. As usual, we first identify all articles that are available in English as well as one of the non-English languages that we address in the shared task. Subsequently, we selected 100 pairs of articles for each language pair, which were later split into two sets, i.e., from and into English.

This year, we aimed to prioritize three topics[5] in our test sets. While selecting the articles, we restricted each topic to around a third of the total. Still, this limit was frequently not reached because too few articles included any of the three selected topics. The three topics are listed below:

- Animals: D000818
- SARS-CoV-2: D000086402
- Pancreatic Neoplasms: D010190

Subsequently, the 100 selected articles for each language pair were split between the two test set directions. Test set statistics are shown in Table 1. No further processing was performed on the test sets, and these were released as a plain text file, one for each language pair, each with 50 lines, and one for each article. Each line is composed of the title and abstract of the article.

3 Baseline/Comparison system

While we used GPT 3.5 as a comparative model last year, we decided to use a self-hosted open-weight large language model this year. Several such models are available of various sizes, licenses, and performance levels in the MT task. Based on the previously accumulated hands-on experience in informally evaluating several open-weight models in multiple tasks, including translation, we selected one of the best performing models, namely Llama 3.1 (Dubey et al., 2024).[6]

The Llama models are open in the sense that their weights and supporting code are freely available, but the usage is limited by a relatively liberal license. In the case of the model used here, the precise licenses are "LLAMA 3.1 COMMUNITY LICENSE AGREEMENT" and "Llama 3.1 Acceptable Use Policy". The last one prohibits using the model to violate the law or the rights of others, to activities related to bodily harm, including military, to generate false information, and includes a clause making it compulsory to report "risky content generated by the model". This risky content can arise when used for medical texts in the form of mistranslated medical procedures.

To interact with the model, we used ollama,[7] through which the model can be queried (i.e. we can programmatically perform tasks with the selected LLM and retrieve the response to those tasks, e.g. from a program written in the Python programming language). In addition, ollama provides a command line interface that can be used to pull further models or to interact with a model in a text-based chat interface.

Implementation decisions We used "Meta Llama 3.1 70B Instruct"[8] (known in ollama as llama3.1:70b), which means the approximate number of parameters is $70 * 10^9$. Such an LLM is run fully accelerated by a GPU only when the parameters fit into the video RAM of the GPU. Since we used a Nvidia A6000 ADA, a 48 GB RAM GPU card, we used the quantization Q4_0 (4 bits per parameter). This makes the actual size of such a model 37.22 GiB and fits in the 48 GB VRAM of the GPU. With the other temporary data needed in the same memory during processing, the occupation of the VRAM went up to 41.2 GB (85%). To evaluate the impact of using the same model with a smaller card, we also tested a 24 GB VRAM card, Nvidia A5000. This raised the CPU usage to 28 cores (from 2) and processing was slower.

[4] https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/

[5] defined as Medical Subject Headings, MeSH terms, used for MEDLINE indexing

[6] https://ai.meta.com/blog/meta-llama-3-1/

[7] https://ollama.com/

[8] "Instruct" indicates that the model was further trained to follow instructions and not just to predict the next text tokens that could follow after a given text prefix.

topics	fr2en	de2en	it2en	pt2en	ru2en	es2en	en2fr	en2de	en2it	en2pt	en2ru	en2es
SARS-CoV-2	15	18	-	5	11	17	9	15	-	-	13	16
Pancr. Neopl.	2	15	1	-	3	2	-	15	2	1	4	1
Animals	15	17	5	17	14	22	20	17	5	18	20	13
other	19	-	44	28	22	9	21	4	43	31	14	20

Table 1: Statistics of the topics in the test sets. The topic "other" refers to articles that do not contain any of the three selected topics. The sum of the values for one language pair might be higher than 50 because some articles contain more than one topic.

Prompt used Choosing the right prompt is important for instruction-tuned LLMs and is still rather an art than a science. We started with the prompt "*You are a helpful assistant specialised in biomedical translation. You will be provided with a text in {src}, and your task is to translate it into {dest}.*" where *src* is the name of the source language and *dest* is the name of the target language.

Visual examination of one text entry (out of the 50 in the test set) per language pair showed the following undesirable behaviour in the MT output generated by the LLM, which we tried to fix by changes to the prompt:

- in one case some additional text, with the meaning "this is the translation into German", which was fixed by adding "*You will add nothing and comment nothing, just produce the accurate translation of the text in specialist language.*" to the prompt;
- additional formatting of the output text through the insertion of newlines, which was almost entirely fixed by adding "*Keep the formatting as close as possible to the source and especially do not insert any newline.*" to the prompt.
- the occasional replacement of digits by their names. We decided not to try to fix this.

After a complete run, we noticed that the LLM still failed to respect the original format of the source texts (it still sometimes produced multiple lines per source text). Visual inspection showed that in a few cases it still attempted to format the subsections of the translated test despite being asked to refrain from doing that. Therefore, explicit postprocessing was carried out to eliminate the line breaks from the LLM's outputs.

Some good features of the translated texts were also noticed, such as localized acronyms e.g. translating English *Real-time functional magnetic resonance imaging (fMRI)* to French *L'imagerie fonctionnelle par résonance magnétique (IRMf)*. Quite impressive was how well the translation retained the quantitative results in the fairly long source texts, while simultaneously applying number localization transformations, such as swapping the decimal point with the decimal comma.

Run-time Statistics Measured duration in seconds with an A6000 in each case for 50 texts:

en2de	1232	en2es	1065	en2fr	1202
de2en	728	es2en	902	fr2en	859
en2it	1413	en2pt	1098	en2ru	1110
it2en	810	pt2en	748	ru2en	641

With an A5000, the speed was about 10 times slower. A GPU-free execution is also possible, but it can be too slow to be practical.

Energy consumption, CO_2 emissions For the A6000 card, a total of $11,607$ seconds at about 1 kW (300W the GPU itself) equals an amount of 3.22 kWh and an equivalent CO_2 emission of 1.16 kg – at the average 360 g CO_2/kWh in Germany, equivalent to the emission of an ICE (internal combustion engine) car driven for about 9.5km. For the slower card, which totalled $131,898$ execution seconds, the figures are 36.64 kWh and therefore 13.2 kg CO_2.

4 Teams and systems

We followed similar dates to the WMT General Translation Shared Task, releasing the test sets on June 27th, 2024 and allowing submission until July 12th, 2024 (after an extension). We released all test sets both in our submission system (Google Form) and the OCELoT tool.[9] We also included our test sets for en2de, en2es, and en2ru as test suites in the General Task[10] in OCELoT. These were the only language pairs that overlapped with the ones from the General Task.

[9] https://ocelot-west-europe.azurewebsites.net/

[10] http://www2.statmt.org/wmt24/translation-task.html

Team ID	Institution	Publication
ADAPT	Dublin City University, Ireland	(Castaldo et al., 2024)
AIST	National Institute of Advanced Industrial Science and Technology, Japan	
DCU	Dublin City University, Ireland	
HW-TSC	Huawei Translation Service Center, China	
Unbabel	Unbabel, Portugal	

Table 2: List of the participating teams and systems.

We received submissions from five teams that directly registered to our task. We list them in Table 2 and present details about their systems below.

ADAPT (Castaldo et al., 2024). For the submissions identified as "run1" for de2en, en2de, fr2en, and en2fr, the participants relied on NLLB-200's distilled 600M variant (NLLB Team et al., 2022), which was fine-tuned on around 10k parallel segments from in-domain training data in the respective language pair. Run2 for en2de, in addition to the above approach, included post-edition by LLM agents powered by GPT-4o.[11] Finally, for run3 for de2en, they relied on LLama-3-8B[12] fine-tuned on around 10k parallel sentences and few-shot prompting using fuzzy matches retrieved by similarity search from the training dataset.

AIST. For run1 of de2en, the team relied on a Mega model (Ma et al., 2023) trained from scratch and fine-tuned on parallel biomedical data from MEDLINE. For run2 for both en2de and de2en, they used a Mega model, an ensemble of four checkpoints trained from scratch and fine-tuned on the same data. For all submissions, they estimate the following sizes of training data used: 3M from in-domain, 5M from open domain, and 3M monolingual.

DCU. We do not have much information about the system behind the submissions for this team, except for a short description citing the Mistra-7B language model[13] for ru2en and fr2en.

HW-TSC. For all submissions to en2de and de2en, the team relied on a system based on Transformers that was trained from scratch on in-domain and open-domain parallel and monolingual data (Wu et al., 2023). It is not clear which changes were carried out for the distinct runs.

Unbabel The submissions for all language pairs consisted of a new version of the Tower LLM (Alves et al., 2024), either with Greedy (run1) or MBR (run2) decoding. The LLM has 70B parameters, was built on top of Llama3, and its continued pre-training phase used 25B tokens for 15 languages, followed by fine-tuning with instructions for all the languages in a variety of tasks, including MT.

5 Automatic evaluation

We ran automatic evaluation based on BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020). We present the results for the submissions to the biomedical translation task using our form in Tables 3 (from English) and 4 (into English), as well as the ones from OCELoT for our task in Table 5 and for our test suites submitted to the General Task in Table 6. All scores were multiplied by 100.

5.1 Biomedical Task submission system

Among all submissions, including our baseline system, the highest BLEU score was 55.63 for pt2en (Unbabel run1) and the highest COMET score was of 89.71 for en2ru (Unbabel run2). The submissions that scored better were the ones from Unbabel and our baseline system, e.g., for en2de, en2fr, en2it, en2pt, and en2ru, with some few exceptions where another system also obtained a high score, e.g., AIST for en2de and DCU for en2ru. The submissions from Unbabel usually scored slightly higher than our baseline, with a few exceptions, e.g., en2pt, fr2en, and es2en.

We observed that the two types of metric score were rather equivalent and that submissions that scored high for BLEU also did so for COMET. However, some submissions had very different BLEU scores for similar COMET scores. For instance, the baseline system obtained the BLEU scores of 31.67 and 51.65 for en2de and en2pt, respectively, but around 87.00 for the COMET score in both cases. Overall, the scores from this year's

[11] https://platform.openai.com/docs/models/gpt-4o

[12] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[13] mistralai/Mistral-7B-v0.1

Team	Run	Metric	en2de	en2fr	en2it	en2pt	en2es	en2ru
ADAPT	1	BLEU	25.03	29.92				
		COMET	84.31	78.14				
ADAPT	2	BLEU	*30.16					
		COMET	85.30					
AIST	2	BLEU	33.80					
		COMET	85.59					
DCU	-	BLEU	16.46		29.12	38.97		31.28
		COMET	64.78		80.39	74.17		87.00
HW-TSC	1	BLEU	*28.77					
		COMET	82.92					
HW-TSC	2	BLEU	28.46					
		COMET	82.83					
HW-TSC	3	BLEU	28.32					
		COMET	83.14					
Unbabel	1	BLEU	34.22	53.54	34.84	50.35		35.76
		COMET	87.48	87.26	85.17	87.03		88.97
Unbabel	2	BLEU	*32.13	*49.76	*32.06	*48.47		*32.35
		COMET	88.09	87.60	86.04	87.55		89.71
Baseline	-	BLEU	31.67	45.98	31.64	51.65	47.95	30.92
		COMET	87.00	87.03	85.00	87.02	85.37	87.55

Table 3: BLEU and COMET scores for submissions to the Biomedical Task submission system, for translation from English. The runs marked with a star (*) were the ones selected for manual validation. For the submissions from Unbabel, runs "1" are the ones identified as "Greedy", and runs "2" are the ones for "MBR".

Team	Run	Metric	de2en	fr2en	it2en	pt2en	es2en	ru2en
ADAPT	1	BLEU	*32.24	18.81				
		COMET	83.04	72.14				
ADAPT	3	BLEU	36.93					
		COMET	78.84					
AIST	1	BLEU	45.86					
		COMET	84.65					
AIST	2	BLEU	*45.92					
		COMET	84.84					
DCU	-	BLEU	32.60	31.47	28.40	31.32	28.02	25.76
		COMET	78.99	78.74	79.63	79.56	80.90	70.01
HW-TSC	1	BLEU	*45.79					
		COMET	83.98					
HW-TSC	2	BLEU	45.68					
		COMET	83.86					
HW-TSC	3	BLEU	45.43					
		COMET	84.08					
Unbabel	1	BLEU	49.05	53.29	38.91	55.63	51.32	47.28
		COMET	86.67	86.05	85.32	85.11	86.99	83.82
Unbabel	2	BLEU	*46.72	*51.67	*38.91	*53.53	*52.28	*45.11
		COMET	86.97	86.39	85.32	85.47	87.25	83.95
Baseline	-	BLEU	45.85	54.79	37.49	51.38	53.54	43.70
		COMET	86.39	86.11	85.28	85.08	87.18	83.37

Table 4: BLEU and COMET scores for submissions to the Biomedical Task submission system, for translation into English. The runs marked with a star (*) were the ones selected for manual validation. For the submissions from Unbabel, runs "1" are the ones identified as "Greedy", and runs "2" are the ones for "MBR".

submissions are not directly comparable to the ones from the previous year since, for the first time, we ran an evaluation on the abstract level.

5.2 OCELoT Biomedical Translation task

Only one team (AIST) submitted to the biomedical task in OCELoT, but also for the same language pairs in our submission system and for our test

Team	Run	Metric	en2de	de2en
AIST	517	BLEU	28.30	
		COMET	83.75	
	542	BLEU		39.68
		COMET		82.55
	544	BLEU		39.68
		COMET		82.55
	545	BLEU	28.30	
		COMET	83.75	

Table 5: BLEU scores for submissions to OCELoT for the Biomedical Translation Task.

suites in the general task. While their results as shown in Table 5 were similar to the ones in Table 6, they were slightly inferior to the ones that the same team obtained for the runs to our submission system, e.g., for en2de, a BLEU score of 28.30 versus 33.80, and a COMET score of 85.59 versus 83.75.

5.3 OCELoT General Machine Translation task

We included test suites only for the language pairs in our task that overlap with the ones considered in the general task, namely, en2de, en2es, and en2ru. The scores for the submissions to the general task (cf. Table 6) varied much more than the ones submitted directly to the biomedical task (cf. Table 3), from very low to high, e.g., BLEU scores of 1.63 (certainly due to mistakes in the system) to 52.56. It is safe to assume that most systems were not trained especially for the biomedical domain. In spite of this, we observed some submissions with scores even higher than the ones for the biomedical task. Amongst the submissions to the general task, the highest scores for en2de were 38.07 BLEU (ONLINE-W) and 88.25 COMET (TransionMT), as opposed to a BLEU score of 34.22 (Unbabel run1) and a COMET score 88.09 (Unbabel run2) in the biomedical task. For en2ru, the highest scores in the general task were 41.25 BLEU (Claude-3.5) and 89.88 COMET (Claude-3.5 and Unbabel-Tower70B), as opposed to 35.76 BLEU (Unbabel run1) and 89.71 COMET (Unbabel run2). Therefore, submissions from the same team (Unbabel) scored slightly higher in the general task than in the biomedical task.

6 Manual evaluation

Similar to previous years, we performed manual validation of a sample of the submissions for most of the language pairs. The number of abstracts that we considered for each language was of either 10 or 20 depending on the availability of the human evaluators. We used the three-way function of the Appraise tool (Federmann, 2018), which includes the following elements:

- the abstract in the original language (e.g., English for en2fr);
- translation A: first translation in the target language (e.g. French for en2fr);
- translation B: second translation in the target language (e.g. French for en2fr).

The task consists of validating whether a translation is better than the other (i.e., A>B or A<B), or whether they are of similar quality (A=B). In cases where the evaluators notice that an error might have occurred, e.g., translation from another text or a translation shorter than it should be, it is possible to skip the validation of this particular pair.

For all language pairs, we considered the best run from each of the team that submitted directly to the biomedical task. The best run was the one identified by the participants during the submission process. Otherwise, we selected the best performing one. We evaluated pairs of either two translations from the teams, or one translation from a team and the reference translation. We present the results for submissions from English in Table 7 and for submission into English in Table 8.

We present below a summary of the mistakes that we observed during manual evaluation.

en2fr Translation quality was uneven, as suggested by the 20 point difference in BLEU scores obtained by the systems. While some translations were of very high quality, others exhibited serious issues including conveying meaning drastically different from the original sentence. In example 1, numerical values are erroneous and inconsistent with the corresponding percentages. In Example 2 the resulting translation is medically unacceptable.

(1) **en:** Of the 273 patients, 164 (60.1%) required invasive mechanical ventilation. One hundred and forty-two patients (52.0%) survived their hospital stay.
fr*: Sur les 273 patients, 104 (60,1%) ont nécessité une ventilation mécanique invasive et 164 (52,0%) ont survécu à leur séjour à l'USI.
fr: Parmi les 273 patient·es, 164 (60,1 %) ont nécessité une ventilation mécanique invasive.

Teams	en2de		en2es		en2ru	
	BLEU	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
AIST-AIRC	28.28	84.85				
Aya23	30.77	87.11	49.49	85.32	31.90	86.69
CUNI-DS					27.93	86.96
CUNI-NL	20.06	83.38				
Claude-3.5	35.23	87.86	52.08	85.93	41.25	89.88
CommandR-plus	32.44	87.67	49.84	85.78	34.33	88.64
CycleL	1.32	38.35	3.00	45.17	0.32	34.65
CycleL2	1.32	38.35			0.10	28.49
Dubformer	31.19	83.49	40.65	78.58	1.94	39.58
GPT-4	35.80	87.93	51.53	85.85	34.00	88.45
IKUN-C	10.82	78.34	22.18	78.23	12.69	81.74
IKUN	11.07	79.14	12.67	74.02	13.28	82.98
IOL_Research	30.86	87.17	48.90	85.56	32.30	87.68
Llama3-70B	31.43	87.01	47.86	85.30	32.18	88.05
MSLC	25.17	82.24	46.30	84.27		
NVIDIA-NeMo	15.91	80.21	30.00	79.32	20.37	83.28
ONLINE-A	36.09	87.34	52.56	85.62	40.20	89.23
ONLINE-B	36.48	88.21	51.56	85.13	40.23	88.73
ONLINE-G	34.86	87.08	50.98	85.34	37.22	89.44
ONLINE-W	38.07	88.04	52.47	85.78	39.77	89.52
Occiglot	6.33	70.19	31.93	78.52		
TSU-HITs	1.63	37.00	17.23	60.20	2.80	52.36
TranssionMT	36.57	88.25	52.67	85.67	40.07	88.76
Unbabel-Tower70B	32.37	87.89	47.93	86.12	32.61	89.88
Yandex					35.09	89.81

Table 6: BLEU scores for submissions to OCELoT for the General Machine Translation Task.

Languages	Systems	Abstracts			
		A>B	**A=B**	**A<B**	**skipped**
en2de	AIST vs. ADAPT	3	3	12	2
	AIST vs. HW-TSC	13	2	4	1
	AIST vs. DCU	10	3	4	3
	AIST vs. reference	2	7	10	1
	AIST vs. Unbabel	2	5	12	1
	ADAPT vs. HW-TSC	16	2	0	2
	ADAPT vs. DCU	10	5	1	4
	ADAPT vs. reference	0	8	10	2
	ADAPT vs. Unbabel	2	10	6	2
	HW-TSC vs. DCU	6	2	9	3
	HW-TSC vs. reference	0	0	19	1
	HW-TSC vs. Unbabel	0	1	18	1
	DCU vs. reference	0	3	14	3
	DCU vs. Unbabel	0	2	15	3
	reference vs. Unbabel	2	10	7	1
en2fr	reference vs. Unbabel	14	0	5	1
	reference vs. ADAPT	17	0	2	1
	Unbabel vs. ADAPT	18	0	1	1
en2it	reference vs. DCU	5	1	13	1
	reference vs. Unbabel	1	1	18	0
	DCU vs. Unbabel	4	6	9	1
en2pt	DCU vs. Unbabel	0	6	8	6
	DCU vs. reference	4	7	3	6
	Unbabel vs. reference	7	10	3	0
en2ru	reference vs. Unbabel	4	2	4	0
	reference vs. DCU	3	3	4	0
	Unbabel vs. DCU	7	2	1	0

Table 7: Pairwise manual evaluation results for the test set (from English).

Languages	Systems	Abstracts A>B	A=B	A<B	skipped
de2en	DCU vs. AIST	3	2	2	3
	DCU vs. Unbabel	2	2	3	3
	DCU vs. reference	2	2	3	3
	DCU vs. HW-TSC	5	0	2	3
	DCU vs. ADAPT	6	0	1	3
	AIST vs. Unbabel	1	0	9	0
	AIST vs. reference	3	2	5	0
	AIST vs. HW-TSC	4	3	3	0
	AIST vs. ADAPT	8	0	2	0
	Unbabel vs. reference	8	2	0	0
	Unbabel vs. HW-TSC	10	0	0	0
	Unbabel vs. ADAPT	10	0	0	0
	reference vs. HW-TSC	6	1	3	0
	reference vs. ADAPT	6	2	2	0
	HW-TSC vs. ADAPT	5	1	4	0
fr2en	DCU vs. ADAPT	6	0	4	0
	DCU vs. reference	1	2	7	0
	DCU vs. Unbabel	0	0	10	0
	ADAPT vs. reference	0	2	8	0
	ADAPT vs. Unbabel	0	0	10	0
	reference vs. Unbabel	1	3	6	0
it2en	reference vs. Unbabel	0	5	15	0
	reference vs. DCU	5	3	8	4
	Unbabel vs. DCU	11	4	1	4
es2en	DCU vs. reference	4	2	9	5
	DCU vs. Unbabel	3	4	8	5
	reference vs. Unbabel	5	6	9	0
ru2en	reference vs. Unbabel	2	2	5	1
	reference vs. DCU	4	0	4	2
	Unbabel vs. DCU	4	3	1	2

Table 8: Pairwise manual evaluation results for the test set (into English).

Cent quarante-deux personnes (52,0 %) ont survécu à leur séjour à l'hôpital.

(2) **en:** Deaths by mechanical asphyxiation constitute a social drama
fr*: La prévention constitue un drame social
fr: Les morts par asphyxies mécaniques constituent un drame social

In both cases, the translation errors likely result from mixing information contained in different parts of the original texts. Arguably, this is very concerning because users of such a translation system could conclude that the erroneous translations are correct by checking that the information is present in the original text. Other issues are more easily detected, such as the interruption of the translation by a loop repetition of a set of tokens (e.g., *une mobilité allant de 5,6% à 5,6% à 5,6% à 1211% à 1211% à 1211% à 1211% à 1211% à 1211% à 1211% à 1211% à 1211%...*).

The choice of having full abstract translation instead of sentence-by-sentence translation this year seems to have both a positive impact on the overall consistency of translations (e.g., overall consistent use of terms and acronyms throughout a document) and a negative impact on the end of translation for some systems, where translation quality was decreasing as the text unfolded and sometimes just interrupted (with or without loop repetitions).

Specialized term translation was sometimes erroneous, in particular with terms referring to animal species (for example, translating *waterfowl* by *oiseaux d'eau* instead of *sauvagine*), which were more frequent this year due to the selection method for the test documents. Polysemous terms were also a source of erroneous translations (e.g., *hood* translated as *capot* – car context instead of *capuche*, which is correct in a clothing context).

In addition to the manual evaluation through appraise, a complementary assessment of the best system submission outputs was conducted, with a focus on *Acronyms* and *Lab Values*, consistently with the evaluations conducted in the two previous years. Overall, 31 out of 50 test documents contained

acronyms and none contained lab values. Acronym translations were considered correct when the system translation was identical to the reference translation or consisted of an attested acronym use in a similar context. Correct acronym translations (79%) included frequent acronyms such as USI (*Unité de Soins Intensifs* – Itensive Care Unit) or IC (*Intervalle de Confiance* – confidence interval). In other cases, acronyms were either untranslated (16%) or erroneous (5%). Some of the acronym translation strategies used by human translators and not by machine translation consist of explicitly stating that an English acronym is used, for example: *la santé mentale du nourrisson (IMH en anglais)*. This is sometimes combined with a strategy of using the long form of a term in French, when an acronym was used in English. These strategies are often used with acronyms that stand for infrequent terms.

It is also interesting to notice that reference translations contain idiomatic linguistic traits not used in machine-translated text, such as inclusive writing (as seen in Example 1).

en2pt All translations into Portuguese were of very good quality, except for some empty translations from one submission and the remains of the prompt used, which were included in the translations of the same submission. Therefore, the decision of whether one translation was better than the other was generally based on small details, often one single mistake.

Small mistakes that we found were the following: (a) lack of capitalization at the start of the sentence (e.g., "…profunda (TVP). o sangue …"); (b) nominal concordance (e.g., "o febre pós-anestésica"); (c) missing words (e.g., "com uma [força] média de 526N"); (d) words that remained in English (e.g., "odds ratio") (e) typos ("registe" instead of "registre"); (f) and grammatical mistakes (e.g., "acompanhou [por] mais de 18 meses").

As in previous years, we found mistakes related to the non-translation of acronyms. For easier or more common terms, e.g., Artificial intelligence (AI), the translations were all correct, i.e., "inteligência artificial (IA)". However, mistakes were often found for other terms, as in Example 3 below in which only the translation pt_3 is correct and has the right acronym:

(3) **en:** Computer vision (CV)
pt_1: visão por computador (CV)
pt_2: visão computacional (CV)
pt_3: visão computacional (VC)

Often we observed a copy of the English acronym for much more complex terms, as in Example 4:

(4) **en:** hydrogenated castor oil (HCO ethoxylates)
pt_1: Óleo de castor hidrogenado polioxietileno (etoxilações de HCO)
pt_2: óleo de rícino hidrogenado de polioxietileno (HCO-etoxilados)
pt_3: hydrogenated castor oil (HCO ethoxylates)

However, we had some difficult examples in which the translation and acronym were correct, e.g. pt_2: in Example 5:

(5) **en:** hospital standardized mortality ratio (HSMR)
pt_1: taxa de mortalidade hospitalar padronizada (HSMR, na sigla em inglês)
pt_2: razão de mortalidade hospitalar padronizada (RMHP)

Finally, we observed many examples in which we favored some translation over others because they either sounded better or more correct, namely, translations pt_2: in Examples 6, 7, 8, and 9:

(6) **en:** was highly expressed in CTCs
pt_1: foi altamente expresso em CTCs
pt_2: tinha uma expressão elevada nas CTCs

(7) **en:** A quasi-experimental study, which compared
pt_1: Estudo quase-experimental, que comparou
pt_2: Um estudo quase experimental, que comparou

(8) **en:** Case signalment
pt_1: Fatores de identificação do caso
pt_2: O sinalamento do caso

(9) **en:** axis of the femoral neck
pt_1: eixo do colo do fêmur
pt_2: eixo do colo femoral

fr2en With the change in protocol this year (from sentence-level to paragraph-level translation and evaluation), there were several differences in the observed quality of translations.

Translation issues brought up in previous years remained present, namely the copying or wrong translation of acronyms and specialised terms (Example 10), the wrong translation of personal pronouns (e.g. *son* 'his/her/their' in Example 11) and errors linked to the ambiguity of source terms (e.g. *taille* 'height or waist' in Example 12).

(10) **fr:** la thérapie de substitution de la nicotine (TSN)
en: nicotine replacement companies (NTS)
en*: nicotine replacement therapy (NRT)

(11) **fr:** . . . la capacité d'un individu à rechercher des soins . . . pour son animal de compagnie
en: an individual's ability to seek . . . care for their companion animal
en*: an individual's ability to seek . . . care for his companion animal

(12) **fr:** la circonférence de la taille (CT)
en: waist circumference (WC)
en*: circumference of height (CT)

However, the overall quality of the translations was visibly lower than in previous years, due to the use of LLMs and the translation of whole paragraphs rather than individual sentences. LLMs tended to exhibit more volatile behaviour, often copying the source document instead of translating, and also including the initial prompt in the output. The consequence of the longer documents to translate was mostly seen in skipping sentences within the documents or (more commonly) at the end of documents (i.e. translation finishing too early or repeating the final sentence multiple times). We also observed the merging of multiple sentences/clauses into a single one and the negative influence of previous sentences on later translations, resulting in the repetition of terms in inappropriate places and errors in the translation of numbers (both problems illustrated in Example 13).

(13) **fr:** Cent six médecins ont répondu au sondage et 12 ont participé à un entretien
en: One hundred and six physicians responded to the survey and 12 participated in an interview
en*: One hundred and twelve respondents participated in the survey

The consequence of the appearance of these more serious errors (i.e. non-translation, missing parts of the translation etc.) meant that they often formed the basis of the evaluation rather than distinctions being based on errors more traditionally resulting from the translation of scientific texts (terminology, acronyms, etc.). Not evaluating on the sentence level meant that an improved translation on the sentence level was easily overridden by a more technical problem, such as a missing sentence at the end of the document. It could be useful in the following years to consider evaluation via error analysis to get more detailed insights into the strengths and weaknesses of different systems on a more granular level.

es2en Contrary to past years, the Spanish to English language pair had very few contributions, totalling 30 examples from two different MT models both compared between each other and against a reference human translation.

In the past, sentence-to-sentence translation has provided good results in terms of translation quality at sentence level. However, the trade-off was inconsistency in the usage of medical terminology and medical specific acronyms. This year however, the use of full abstracts for translation led to greater consistency in the translation of terminology and acronyms specific to medicine.

When working well, the MT output has a good quality, sometimes producing a result that was comparable to human translation in terms of quality, as shown in Table 8, where the MT system Unbabel had very good results compared against DCU and the reference translation.

However, the MT output still lacks the fluency of a human translation, as the systems had a tendency to replicate the structure of the original Spanish source text, resulting in translations that can be considered "literal translations". In many instances, the MT output would require copy editing and rewriting by a native English speaker to render the text more fluent and increase the overall quality of the output.

Despite the good quality level of some translations, the overall quality of the outputs for this year's challenge is very uneven, with some very good abstracts in English and some abstracts that were not translated or still contained Spanish words in them.

At least one of the system used LLMs to produce the output in English, with this prompt: "1. While being factual, accurate and not missing out any detail, translate the given Spanish text into the specified English language. Spanish Text:". The use of the prompt ensured the output did not miss information from the original source text, as has sometimes been the case in past years. Nevertheless, the LLM system was not very robust.

As shown in the example below, the LLM system sometimes did not translate the text in English as requested. The text remained in Spanish. That is considered a missing translation and is considered a major error.

(14) **en:** While being factual, accurate and not missing out any detail, translate the given Spanish text into the specified English language. La prevalencia de alergia alimentaria ha aumentado en algunas regiones del mundo, y con ello la incidencia, según la variabilidad geográfica, en el fenotipo y manifestaciones clínicas...

Another error the LLM system made was the inclusion of the prompt used to generate the translated output as part of the response. This add superfluous information to the English translation and breaks the readability and fluency of the text (see previous example).

As mentioned before, fluent translation was still an issue for the machine translation system, in particular for the DCU system. This system sometimes generated sentences that were clunky or ungrammatical in English.

(15) **es:** Se registraron 4 casos de morbilidad post punción (2 dolores epigástricos y 2 hematomas de pared abdominal

en: Were registered 4 cases of morbidity post puncture (2 pain epigastric and 2 hematomas of abdominal wall).

In conclusion, LLMs systems still seem to have an unreliable performance when it comes to machine translation, producing very good quality translations, missing translations or ungrammatical translations at the same time. A better out-of-box LLM or refine the prompting techniques might obtain better results with these systems.

It must be noted, however, that there were very few examples for the Spanish to English translation to reach an indisputable conclusion.

en2de Similar to previous years, a generally high level of translation quality was seen for English to German translation. The strongest models produced translations that not only conveyed the content well but also maintained consistency in terms of style and structure. However, certain systems exhibited notable flaws. In particular, one model consistently omitted portions of the text, often truncating the translation towards the end of the document or, at times, even mid-sentence. Another system struggled with basic capitalization, failing to begin sentences with an uppercase letter, which detracted from the overall readability of the output.

Numerical translations were also an issue, with *Eighty-nine* frequently mistranslated as either *Achtundachtzig* "eighty-eight" or *Achtundneunzig* "ninety-eight", revealing inaccuracies in number handling. The translation of abbreviations varied across systems, with some attempting to expand or translate them, occasionally resulting in errors. For example, the *European Commission (EC)* was incorrectly translated as *EG (Europäische Gemeinschaft)* instead of EU. Furthermore, specialized terminology presented additional challenges, with terms like *compulsory elective* rendered awkwardly as *obligatorische elektive Veranstaltung* rather than the more appropriate *Wahlpflichtkurs*.

Grammatical errors also persisted in some translations, indicating that while overall quality was high, there is still room for improvement in handling both sentence structure and more nuanced linguistic elements.

de2en Overall, results varied for the German-to-English translation task. While at least one system was able to provide a human-level translation for each source sample, there was generally also at least one translation that was either incomplete or difficult to understand.

The most serious mistakes included omission of whole sentences, or synthesis of text that was not present in the original. This was especially evident in cases where the sample text ended in an incomplete sentence, which caused some systems to generate a completion to the sentence. In the most egregious example of this phenomenon, an incomplete sentence at the end of a description of an animal's skin condition after an insect bite led to more than one translation mentioning euthanasia, when no such language was present in the source. In some instances, text would be translated to nonexistent words, e.g. translation of *porös* to

the nonexistent word *sporeous*. Other mistranslations included rendering *mittleren Werte* as *median* instead of *mean* values as was intended in the text.

The most frequently occurring mistakes were related to the capitalization of words at the beginning of sentences. Other formatting mistakes failed to take into account the structure of the text, omitting paragraph headings. These mistakes did not affect the overall intelligibility of the text.

All in all, the majority of the systems were able to provide a translation that, while not perfect, was understandable and correctly conveyed important information.

en2it The quality of the translation was higher than in previous years, even more so than last year, which set a new threshold in the accuracy of the translation from English to Italian and vice versa. The quality of most of the abstract was almost identical and fluent in terms of the quality of language. The terminology and the syntax was of very high quality in both translation directions. There were rarely major issues with the choice of terms or the construction of the sentences.

One mistake was the addition of parts of the text that were not present in the original version. For example, the original version is “Among those diagnosed with COVID-19 during follow-ups between March 2020 and March 2021 [...]”

While the Italian translation: “MATERIALE E METODO: TRA marzo 2020 e maggio 2021, sono stati analizzati [...]”

Where there is the addition of “MATERIALE E METODO”. There is also some minor issue with the punctuation (the semicolon between “rene” and “o dobbiamo farlo” should not be there) as well as uppercase letters (“TRA” instead of “tra”).

There were two problems concerning the cause effect or correlation among pathologies. For example, in the original English version: "Chronic rhinosinusitis with nasal polyps is a common disease with still unclear pathophysiologic mechanisms." The "Chronic rhinosinusitis with nasal polyps" are one thing all together that is documented to be a common disease.

On the other hand, the Italian version: "La rinosinusite cronica e la poliposi nasale sono patologie frequenti" the "Rinusite cronica" ("Chronic rhinosinusitis) and "poliposi nasale" ("nasal polyps") are considered as two distinct pathologies.

The other example happens with the following sentence: "The airway epithelial barrier has been shown to be involved in different chronic disorders, including rhinitis, nasal polyposis and asthma" and its Italian translation: "La barriera epiteliale delle vie respiratorie sembra essere coinvolta in diverse patologie croniche come la rinite, la poliposi nasale e l’asma"

In this case, the translation gives a slightly different interpretation of the fact that, in the original version, "airway epithelial barrier has been shown to be [...]" as in "it has been demonstrated that", while the Italian "sembra essere coinvolta" ("seems to be involved") shoes a less strong connection between the entities (airway epithelial barrier and chronic disorders).

it2en For the Italian to English translation direction, we observe an opposite problem compared to the English one that is removing a part of the text.

For example, in the original "Conclusione: sebbene non abbiamo riscontrato differenze significative tra i pazienti sottoposti a gastrectomia standard e quelli sottoposti a NACT prima della gastrectomia, [...]" we have "Conlusione:" as the initial part of this sentence.

In the English version, we have "Although we found no significant difference between the patients undergoing standard gastrectomy and those undergoing NACT before gastrectomy," Where "Conclusions" ("conclusione") is missing.

From Italian to English, there was a missing agreement in gender for the translation of the following sentence: "A total of 192 female feral cats were investigated for a large-scale trap-neuter-release program." One of the Italian translations overlooked the female gender with: "Un totale di 192 gatti selvatici sono stati studiati per un ampio programma di trappola, sterilizzazione e rilascio." Where "gatti" is the masculine plural of a cat which, in this case, is wrong.

Another type of wrong accordance was found in the translation of the following sentence: "La gangrena di Fournier è una fascite necrotizzante a rapida progressione che coinvolge il perineo, le regioni perianale e genitali e costituisce una vera emergenza chirurgica con un tasso di mortalità potenzialmente elevato" where the English version: "Fournier’s gangrene is a rapidly progressing necrotizing fasciitis involving the perineal, perianal, or genital regions and constitutes a true surgical emergency with a potentially high mortality rate." considers the "perineal [...] region" instead of the "perineum" alone.

en2ru and ru2en This year, two systems, Unbabel and DCU, participated in the Biomedical Machine Translation task. Generally, the translations to and from English were of high quality. We did not encounter examples that were completely unacceptable, aside from a few cases where text boundaries were mapped incorrectly. Compared to previous years, we observed a general improvement in how the systems handled abbreviations, which is a notable challenge in biomedical translation.

This year translations were evaluated at the abstract level, and at times determining which translation was superior often came down to small details. In some instances, we preferred one translation over another purely due to stylistic differences. There were only a handful of cases where the systems diverged significantly in quality. Overall, Unbabel outperformed DCU, as reflected by manual evaluation (Table 7 and 8) and better BLEU and COMET scores (Tables 3 and 4).

7 Conclusions

We presented the results for this year's edition of the Biomedical Translation Task at WMT, in which we considered 12 language pairs. In this paper, we described the development of the test sets, the submissions we received, our baseline system, and the details about the automatic and manual evaluation. Different from previous years, we did not split and align the sentences, instead we had the test sets simply composed of the title and abstracts of the articles.

Limitations

Concerning the quality of the extracted test sets, the passage from sentence to paragraph level is likely to require additional post-processing in future years. Whereas in previous years, sentence alignment resulted in additional validation of the extraction process, a number of errors were present in the test sets this year, resulting in more skipped evaluations. These included (i) missing or additional sentences in the reference translations with respect to the source texts, (ii) the truncation of certain sentences after special characters and subscript text, the inconsistent inclusion of headers (e.g. *Methods*, *Results*) in the abstracts and the non-capitalised of accented characters in the headers (e.g. French *RéSUMé* 'Abstract' instead of *RÉSUMÉ*), a consequence of the original source text, but which could be corrected in a post-processing step.

Ethics Statement

Our test sets were derived from PubMed, a database of biomedical citations. These publications are used in many areas of medicine, including decisions about the diagnosis and treatment of patients. Machine translation in this domain should be used as part of a larger framework that should include human experts for the interpretation of translations and, if necessary, the correction and adaptation of the generated text.

Acknowledgments

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR under the project MaTOS - "ANR-22-CE23-0033-03" and as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001. Lana Yeganova's work was supported by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health of the USA.

References

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.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Antonio Castaldo, Maria Zafar, Prashanth Nayak, Rejwanul Haque, Andy Way, and Johanna Monti. 2024. The SETU-ADAPT Submission for WMT 24 Biomedical Shared Task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, Miami, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór 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*, USA. Association for Computational Linguistics.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. *Preprint*, arXiv:2209.10655.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 biomedical translation shared task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 324–339, Belgium, Brussels. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farre-maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff

Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

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.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. The path to continuous domain adaptation improvements by HW-TSC for the WMT23 biomedical translation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 271–274, Singapore. Association for Computational Linguistics.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.