The SETU-ADAPT Submissions for WMT24 Biomedical Shared Task

Antonio Castaldo[a*], **Maria Zafar**[*], **Prashanth Nayak**[b],
Rejwanul Haque, **Andy Way**[c], **Johanna Monti**[a]
[a]UNIOR NLP Research Group, University of Naples "L'Orientale", Naples, Italy
South East Technological University, Carlow, Ireland
[b]KantanAI, Dublin, Ireland
[c]ADAPT Centre, Dublin City University, Dublin, Ireland
antonio.castaldo@phd.unipi.it,c00304029@setu.ie,pnayak@kantanai.io
rejwanul.haque@setu.ie,andy.way@adaptcentre.ie,jmonti@unior.it

Abstract

This paper presents SETU-ADAPT's submissions to the WMT 2024 Biomedical Shared Task, where we participated for the language pairs English-to-French and English-to-German. Our approach focused on fine-tuning Large Language Models (LLMs), using in-domain and synthetic data, employing different data retrieval strategies. We introduce a novel MT framework, involving three autonomous agents: a Translator Agent, an Evaluator Agent and a Reviewer Agent. We present our findings and report the quality of the outputs.

1 Introduction

Translating texts in the biomedical domain presents unique challenges that sets it apart from general domain translation tasks. The domain is characterised by the use of specialised terminology, fixed expressions and relative data scarcity. In recent times, LLMs (Brown et al., 2020; Przystupa and Abdul-Mageed, 2019) have become the go-to systems for building Machine Translation (MT) systems, due to their impressive performance in generating accurate translations across diverse domains. Precisely, the ability to fine-tune these models on new data, adapting them to the specialised terminology used in the biomedical domains, makes them particularly suitable for our task.

In our experiments, we built our MT systems using Llama-3 (Dubey et al., 2024) and No Language Left Behind (NLLB) (Costa-jussà et al., 2022), based on the high performance reported in their relevant publications. We further design and develop strategies to address data scarcity and improve the quality of the outputs. Our first approach involves back-translation (Xu et al., 2019), where we leverage monolingual data and translate them back into the source language, thus generating synthetic data to be combined with the original dataset. This approach is widely recognized as an effective method to overcome the challenges caused by the translation of low-resource languages and specific domains. Another data augmentation method that we adopt is based on terminology-aware mining (Haque et al., 2020), where we extract a terminology list from our training data and use it to mine semantically similar sentences from the general domain corpus. We further experimented using few-shot prompting, where we provided the model with a few translation samples retrieved through semantic search based on the source sentence. Finally, we propose an innovative MT system powered by `GPT-4o` (OpenAI et al., 2024) that employs an agentic workflow (Wang et al., 2024). This system follows a collaborative framework, where three LLM-based agents work together autonomously to produce translations.

The paper is organised as follows. We present an overview of our proposed systems in Section 3. We describe our datasets and our data augmentation strategies in Section 4 and Section 5. We introduce our last system, involving LLM-based autonomous agents in Section 6. We present the results of our evaluation in Section 7 and draw our conclusions in Section 8.

2 Related Work

The main difficulties found in biomedical MT have been the highly specialised domain, the lack of relevant data, and the importance of using the correct terminology. To address the issues caused by

[*]Both authors are equal contributors to this work.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 647–653
November 15–16, 2024.

domain-specific terminology, Choi et al. (2022) adopted a soft-constrained translation approach, where terminology constraints retrieved from the training corpus are provided to the MT system as a suggestion rather than a hard constraint. Soft-constrained decoding appears to be a promising solution to drive the systems to include the necessary terminology in the output while preserving the model's fluency and flexibility in the translation.

Ballier et al. (2022) trained different systems on a selection of texts from WMT, Khresmoi (Dušek et al., 2017) and UFAL (Bojar et al., 2017) datasets, comparing the results. Interestingly, they find that `mBART-50` (Tang et al., 2021), despite producing fluent grammatical sentences, fails at translating consistently domain-specific terminology. Their study suggests that this well-known model may not be adequate for the task of biomedical translation, especially in the context of translating biomedical abstracts where its small context window may cause inaccurate translations.

Manchanda and Bhagwat (2022) confirms previous studies that showed how fine-tuning any model from a general domain to a specialised one, as is the case with clinical and biomedical texts, improves the translation quality in most cases. Their study introduces a novel approach, based on combining general-purpose and domain-specific datasets for fine-tuning while applying a higher learning rate to the general domain data. Their experiments demonstrate how this combined fine-tuning approach may improve translation quality in both domains.

In the last few years, we have seen a general surge of LLMs applied to MT (Hendy et al., 2023; He et al., 2024). Several studies have been conducted with a high degree of success on the application of LLMs for the translation of biomedical texts. The first study of this sort was published by Han et al. (2022), where they compare MT models of different sizes to investigate the applicability of Kaplan's scaling laws (Kaplan et al., 2020) in biomedical translation. Their findings confirmed that larger general-purpose models consistently outperform smaller models, even when the latter are fine-tuned on domain-specific data. Interestingly, the performance gap narrows significantly when the training data for smaller models is meticulously curated, bringing their efficacy close to that of the NLLB model. The efficacy of LLMs in translating biomedical data is further confirmed by several recent studies (Jahan et al., 2024; Keles et al., 2024; García-Ferrero et al., 2024).

Finally, we underline one of the latest research directions in the study of LLMs: multi-agentic workflows. Agents are instances of LLMs, each with a tailored system prompt that defines their behaviour, adhering to specific criteria and output requirements. Usually, agents also have access to external features, such as memory mechanism (Zhang et al., 2024), retrieval-augmented generation (Gao et al., 2024), and tool use (Qu et al., 2024). In the experiment conducted by Liang et al. (2024), the authors exemplify this approach with a novel translation framework called Multi-Agent Debate (MAD). Their system is based on a guided interaction between multiple agents who engage in a debate to determine the most effective translation for a given source text. A designated judge agent oversees this process and ultimately decides on the final solution. This iterative strategy allows successive agents to refine the initial translation hypothesis, progressively improving translation quality. They achieve good performance with the models `gpt-3.5-turbo` and `gpt-4` (Brown et al., 2020).

3 Systems Overview

We submit five MT systems for evaluation, each employing different approaches to biomedical translation. These systems range from traditional fine-tuning on in-domain data to various data augmentation approaches and the use of LLMs, prompt engineering, and multi-agent workflows.

Table 1 provides an overview of the five systems submitted for evaluation. System 1 utilizes the NLLB model fine-tuned with terminology mining techniques, applied in both directions (see §5.1). System 2 also uses NLLB, but we fine-tune it on both in-domain and synthetic data. For this system, we augmented the training data with an additional 5,000 backtranslated sentences to address data scarcity (see §5.2). System 3 uses a combination of agents powered by NLLB and GPT, who are tasked with post-editing and refining the NLLB outputs to make them more fluent and effective. For System 4, we select the smallest checkpoint of the most recent models developed by Meta AI, called `LLama-3-8B` (Dubey et al., 2024). This system uses parameter-efficient fine-tuning on the in-domain data, and the output is improved with three fuzzy matches prepended to the prompt. Finally, our last submission, System 5 uses a multi-agent crew powered by `GPT4-o` (OpenAI et al., 2024).

System	Model	In-Domain FT	Backtranslation	Terminology Mining	Agents	FSP
1	NLLB	✓		✓		
2	NLLB	✓	✓			
3	NLLB	✓			✓	
4	LLama-3	✓				✓
5	GPT-4o				✓	✓

Table 1: Overview of our submitted systems. The checkmark (✓) indicates the presence of a feature. FSP stands for Few-Shot Prompting.

The multi-agents workflow is described in depth in the relevant section (see §6).

4 Dataset Selection

In this section, we describe the composition of the datasets used for our experiments. We curated a selection of parallel sentences from the corpora provided by the shared task organisers, including part of the Biomedical Translation repository and the UFAL Medical[1] corpus. This resulted in two datasets: 11,190 parallel sentences for English-German and 13,032 for English-French. We investigated synthetically increasing the training data by employing different data augmentation techniques for the English-to-German language pair. We provide an overview of the dataset selection in Table 2.

Dataset	EN-DE	EN-FR
Original	11,190	13,032
+ Term. Mining	14,583	NA
+ Backtranslation	16,190	NA

Table 2: Overview of datasets.

5 Data Augmentation

In this section, we describe the different approaches we have used to augment the datasets used for our MT systems. We adopt back-translation, terminology mining, and fuzzy matches.

5.1 Terminology Mining

We perform terminology mining on English-to-German language pairs. We extract biomedical terms from the training data using the pre-trained named entity recognition (NER) model `d4data/biomedical-ner-all`. This model is designed to identify biomedical entities within the text, such as diseases, disorders, and therapeutic procedures, providing a confidence score and the specific unit being identified. The implementation utilises the pipeline function from the Hugging Face Transformers[2] library (Wolf et al., 2020), configured for the task of token classification.

First, the NER model iterates over every term in the dataset, obtaining a list of identified entities. We then filter them, collecting only those labeled as *B-Disease-disorder* or *B-Therapeutic-procedure*, provided that the model's confidence score for the entity exceeds 0.98 and the length of the identified word is greater than five characters. Entities meeting these criteria are then printed for verification and appended to the list of extracted terms. This approach ensures that only relevant biomedical terms are extracted from the dataset, focusing specifically on diseases, disorders, and therapeutic procedures. The terminology mining process yielded a total of 14 biomedical terms, that we used to collect 3,393 sentences containing at least one biomedical term.

5.2 Backtranslation

We adopt back-translation for the English-German language pair, to address data scarcity with semantically similar sentences, extracted with a pre-trained sentence embedding model, and then backtranslated with NLLB. We initially filter the EMEA monolingual dataset (Calzolari et al., 2012), selecting only sentences that exceed 100 characters in length, and further limit our selection to the first 1,000 entries for easier processing. We encode the text data using the pre-trained sentence embedding model `multi-qa-mpnet-base-dot-v1` from the Sentence-Transformers library. The sentence embeddings are stored as a new column in the dataset, on which we perform semantic search, using FAISS index (Douze et al., 2024) for more efficient computation. The index is then queried to retrieve the top 5 most similar samples from the original dataset. We collect a total of 5,000 sentences, aggregated

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus

[2] https://github.com/huggingface/transformers

and sorted by similarity scores in descending order.

This methodology enables the efficient retrieval of contextually relevant sentences from large datasets. We make use of the resulting sentences, backtranslating them to English using the baseline `NLLB-200-600M` and leading to the creation of a synthetic dataset, that is in the same domain. The synthetic dataset is added to the original dataset to fine-tune the baseline model.

5.3 Fuzzy Matches

Fuzzy matches are human translated segments, stored in parallel datasets. Drawing on findings from Moslem et al. (2023), we incorporate semantically-similar fuzzy matches in a three-shot prompting scenario. This approach leverages the model's in-context learning ability (Brown et al., 2020) to further improve the quality of the MT outputs. A wide range of academic literature has demonstrated that incorporating fuzzy matches in a few-shot scenario may improve the model's understanding of domain-specific terminology and fixed expressions (Castaldo and Monti, 2024; Moslem et al., 2022; Knowles et al., 2018).

To extract fuzzy matches, we employ semantic search on sentence embeddings generated by the `all-MiniLM-L6-v2` model. The embeddings are stored in a flat index created with the FAISS[3] library, from which we retrieve the three most similar sentences. After extracting the fuzzy matches for our input sentence, we prepend them to a minimalist prompt that directly maps the source language to the target language. We incorporate fuzzy matches in System 4 and 5, achieving substantial improvements over the zero-shot baseline. We present an overview of the prompt templates used in this study in Table 3, with the following annotations: ♦ shows the presence of a line break, `[src]` stands for source language, `[tgt]` stands for target language, and `[input]` stands for the text to be translated.

Prompt Type	Template
Zero-Shot	[src]: [input] ♦ [tgt]:
Few-Shots	[src]: [source$_1$] ♦ [tgt]: [target$_1$] ♦ ... [src]: [source$_k$] ♦ [tgt]: [target$_k$] ♦ [src]: [input] ♦ [tgt]:

Table 3: Overview of the prompt templates used in this study.

[3] https://github.com/facebookresearch/faiss

6 Multi-Agents Workflow

We design a team composed of three autonomous agents that collaborate to simulate a translation agency with the goal of refining an initial translation hypothesis from multiple perspectives. The process begins with the creation of our agent crew, using the CrewAI library[4].

The first agent, the Translator Agent, is tasked with translating a given sentence. Following this, the Evaluator Agent assesses the translation based on fluency and accuracy. This assessment is quantified with a numerical quality metric that ranges from 0 to 100, where 100 signifies a translation that is both perfectly fluent and accurate.

If the translation receives a score below 80, the Reviewer Agent intervenes to review the initial hypothesis, aiming to improve its accuracy. This iterative process repeats until the Evaluator Agent awards a quality score greater than 80, indicating a successful translation. We provide the reference code used for this experiment in the relevant GitHub repository.[5]

7 Evaluation

This section discusses the results that we obtained from our experiments. Table 4 shows the results obtained by evaluating our models on the validation set, using BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and the COMET model `wmt22-comet-da` (Rei et al., 2022a). Our quality estimation is based on the reference-free COMET model `wmt22-cometkiwi-da` (Rei et al., 2022b). Our systems achieve good results for both language pairs, and the data augmentation approaches visibly improve the translation outputs, as documented in the evaluation of Systems 1, 2 and 3. Terminology mining seems particularly effective, improving the BLEU score of our first system significantly above the others.

In order to confirm the results of our automatic evaluation and to allow for a more precise comparison of the different systems used for the primary language pair of our study, we include a manual evaluation on a small sample of translations, conducted by two professional translators in the EN-DE language pair, for which we adopt the MQM-DQF framework (Burchardt, 2013; Lommel and

[4] https://github.com/crewAIInc/crewAI

[5] https://github.com/Ancastal/biomedical-wmt-agents

Melby, 2018).

The results of the MQM evaluation reveal that System 5 produces the fewest overall errors, with the majority of these errors falling under the Fluency category. In contrast, the other systems exhibit a higher concentration of errors in the Accuracy category. In System 1, where Terminology Mining was applied, fewer terminology-related errors were detected, further confirming the effectiveness of this strategy. Additionally, we find that in System 3 the involvement of `GPT-4o` agents in post-editing led to a reduction in accuracy-related errors.

System	BLEU	ChrF	COMET	QE
English-to-German				
Baseline	2.97	26.01	70.09	0.62
System 1	25.29	60.13	79.50	0.58
System 2	23.80	58.89	78.33	0.56
System 3	23.97	59.22	78.93	0.63
System 4	22.95	58.90	84.32	0.73
System 5	25.24	63.01	86.13	0.77
English-to-French				
System 3	29.18	57.01	75.70	0.65

Table 4: Experiment Results for Different Systems

8 Conclusions

This study presents the approaches we have adopted to address the challenges caused by biomedical translation, specifically the need for consistent translation of domain-specific terminology and the lack of in-domain parallel data.

By adopting data augmentation techniques, we found that our models improved consistently in translating biomedical terminology, achieving better results in our evaluation. Terminology mining proved particularly effective, resulting in our best overall submission. We also explored the use of backtranslation, but we found that its effectiveness may be limited in fine-tuning LLMs. We speculate that it may require a different ratio of original to synthetic data used during training, or a different weighting. Our experiments with fuzzy matches demonstrated the potential to use in-context learning to improve MT quality and adapt LLMs to domain-specific terminology.

Finally, we introduced a novel MT workflow based on the collaboration of three autonomous LLM-based agents. This approach offers an innovative way to refine an initial translation hypothesis from multiple perspectives, potentially leading to more accurate outputs.

9 Limitations

We acknowledge that several aspects of our study have room for improvement. First, the evaluation was conducted on a relatively small dataset of 50 biomedical abstracts, limiting the objectivity of the results. Second, while data augmentation helped improve performance, the training data could be expanded by incorporating larger corpora and potentially leading to better quality. Additionally, the models employed in this study may not represent the best performing MT systems by the time of publication, requiring further experiments with more recent models to validate our findings. Finally, manual evaluation was only conducted for a single language pair, limiting the scope of our analysis.

10 Acknowledgements

Part of this work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples “L’Orientale", through a doctoral grant established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

References

Ballier, N., Yunès, J.-b., Wisniewski, G., Zhu, L. (2022). The SPECTRANS System Description for the WMT22 Biomedical TaskIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 895–900, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Bojar, O., Haddow, B., Marecek, D., Sudarikov, R., Tamchyna, A. (2017). Report on building translation systems for public health domain. UFAL Medical Corpus.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS ’20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Burchardt, A. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. (2012). Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12).

Castaldo, A. (2024). Prompting Large Language Models for Idiomatic Translation. In *Proceedings of the First Workshop on Creative-text Translation and Technology*, pages 37–44, Sheffield, UK. Accepted: 2024-06-19T21:00:05Z.

Choi, Y., Shin, J., Ryu, Y. (2022). SRT's Neural Machine Translation System for WMT22 Biomedical Translation TaskIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 901–907, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J. et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672 [cs].

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L. (2024). The Faiss library. arXiv:2401.08281 [cs].

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783 [cs].

Dušek, O., Hajič, J., Hlaváčová, J., Libovický, J., Pecina, P., Tamchyna, A. (2017). Khresmoi summary translation test data 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs].

García-Ferrero, I., Agerri, R., Atutxa Salazar, A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramirez-Romero, J., Rigau, G. et al. (2024). MedMT5: An Open-Source Multilingual Text-to-Text LLM for the Medical DomainIn Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177, Torino, Italia. ELRA and ICCL.

Han, L., Erofeev, G., Sorokina, I., Gladkoff, S. (2022). Examining Large Pre-Trained Language Models for Machine Translation: What You Don't Know about ItIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Haque, R., Moslem, Y. (2020). Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT's Submission to the Adap-MT 2020 English-to-Hindi AI Translation Shared TaskIn Sharma, D. M., Ekbal, A., Arora, K., Naskar, S. K., Ganguly, D., L, S., Mamidi, R., Arora, S., Mishra, P., editors, *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S. (2024). Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv: 2302.09210.

Jahan, I., Laskar, M. T. R., Peng, C. (2024). A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361 [cs, stat].

Keles, B., Gunay, M. (2024). LLMs-in-the-loop Part-1: Expert Small AI Models for Bio-Medical Text Translation. arXiv:2407.12126 [cs].

Knowles, R., Ortega, J. (2018). A Comparison of Machine Translation Paradigms for Use in Black-Box Fuzzy-Match RepairIn Astudillo, R., Graça, J., editors, *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 249–255, Boston, MA. Association for Machine Translation in the Americas.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, R., Yang, Y., Tu, Z. (2024). Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. arXiv:2305.19118 [cs].

Lommel, A. (2018). Tutorial: MQM-DQF: A Good Marriage (Translation Quality for the 21st Century)In Campbell, J., Yanishevsky, A., Doyon, J., editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, Boston, MA. Association for Machine Translation in the Americas.

Manchanda, S. (2022). Optum's Submission to WMT22 Biomedical Translation TasksIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R.,

Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 925–929, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Moslem, Y., Haque, R., Kelleher, J. (2022). Domain-Specific Text Generation for Machine TranslationIn Duh, K., editors, *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Moslem, Y., Haque, R., Kelleher, J. D. (2023). Adaptive Machine Translation with Large Language ModelsIn Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N. et al., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S. et al. (2024). GPT-4 Technical Report. arXiv:2303.08774 [cs].

Papineni, K., Roukos, S., Ward, T. (2002). Bleu: a Method for Automatic Evaluation of Machine TranslationIn Isabelle, P., Charniak, E., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluationIn Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Przystupa, M. (2019). Neural Machine Translation of Low-Resource and Similar Languages with Back-translationIn Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A. et al., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Qu, C., Dai, S., Wei, X., Cai, H., Wang, S., Yin, D., Xu, J. (2024). Tool Learning with Large Language Models: A Survey. arXiv:2405.17935 [cs].

Rei, R., C. de Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L. (2022a). COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared TaskIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L. et al. (2022b). CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared TaskIn Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M. et al., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J. (2021). Multilingual Translation from Denoising Pre-TrainingIn Zong, C., Xia, F., Li, W., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y. et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. et al. (2020). HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs].

Xu, N., Li, Y., Xu, C., Li, Y., Li, B., Xiao, T. (2019). Analysis of Back-Translation Methods for Low-Resource Neural Machine TranslationIn Tang, J., Kan, M.-Y., Zhao, D., Li, S., editors, *Natural Language Processing and Chinese Computing*, pages 466–475, Cham. Springer International Publishing.

Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z. (2024). A Survey on the Memory Mechanism of Large Language Model based Agents. arXiv:2404.13501 [cs].