Occiglot at WMT24: European Open-source Large Language Models evaluated on Translation

Eleftherios Avramidis[(1)], **Annika Grützner-Zahn**[(1)], **Manuel Brack**[(1,2)],
Patrick Schramowski[(1,2,3)], **Pedro Ortiz Suarez**[(4)], **Malte Ostendorff**[(5)], **Fabio Barth**[(1)],
Shushen Manakhimova[(1)], **Vivien Macketanz**[(1)], **Georg Rehm**[(1)], **Kristian Kersting**[(1,2,3,6)]
[(1)]German Research Center for Artificial Intelligence (DFKI), Germany
[(2)]Computer Science Department, TU Darmstadt [(3)]Hessian.AI [(4)]Common Crawl Foundation
[(5)]Occiglot [(6)]Centre for Cognitive Science, TU Darmstadt

Abstract

This document describes the submission of the very first version of the Occiglot open-source large language model to the General MT Shared Task of the 9th Conference of Machine Translation (WMT24). Occiglot is an open-source, community-based LLM based on Mistral-7B, which went through language-specific continual pre-training and subsequent instruction tuning, including instructions relevant to machine translation. We examine the automatic metric scores for translating the WMT24 test set and provide a detailed linguistically-motivated analysis. Despite Occiglot performing worse than many of the other system submissions, we observe that it performs better than Mistral7B, which has been based upon, which indicates the positive effect of the language specific continual-pretraining and instruction tuning. We see the submission of this very early version of the model as a motivation to unite community forces and pursue future LLM research on the translation task.

1 Introduction

Occiglot, initiated in March 2024, is a community-based open-source initiative for "Polyglot Language Models for the Occident". We believe that our dedicated language modeling solutions will not only maintain Europe's academic and economic competitiveness and AI sovereignty, but also have a profound Impact on the preservation of linguistic diversity, multilingualism, and cultural richness. Occiglot is an academic, non-profit research collective committed to open science and open-source LLM development.

Although Occiglot is in the early stages of development, it entails a significant amount of work for large-scale data collection, model pre-training and tuning, and multi-faceted evaluation. Since LLMs can be used in various use cases, targeted evaluation, starting in the first stages, is important for revealing strengths and weaknesses. The shared task of the 9th Conference of Machine Translation (WMT24; Kocmi et al., 2024a) provides the opportunity for testing the performance of the LLM in a translation task.

First, this paper reviews some indicative items of related work (section 2). Then, in section 3 we present the details on the development of the Occiglot model (section 3.1), the training data related to translation (section 3.2) and the engineering towards machine translation and outline the issues and directions for further improvements. Section 4 presents the evaluation, whereas a conclusion is given in section 5.

2 Related work

Prompting LLMs for translation output has been successfully employed since the early years of LLMs (Brown et al., 2020), with the few-shot enhanced context approach indicating good results (Vilar et al., 2023). Later approaches suggested that an adaptive method of few-shot prompting may be even more beneficial (Agrawal et al., 2023; Zhang et al., 2023; Soudi et al., 2024). Enis and Hopkins (2024) deal with evaluating Claude 3 Opus, as compared to other LLMs, with regard to machine translation of low resource languages.

The motivation of Occiglot, to focus LLM development on languages other than English, is confirmed by Diandaru et al. (2024), who suggest that models centered around languages other than English could provide a more efficient foundation for multilingual applications. Zan et al. (2024) follow a similar approach to ours, including instruction tuning tailored to particular target languages. Stap et al. (2024) suggest that including monolingual data as part of the fine-tuning data, we can maintain the abilities while simultaneously enhancing overall translation quality.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 292–298
November 15–16, 2024.

3 The language model

3.1 Training

The submission at WMT24 is based on the current, first version (v0.1) of the Occiglot bilingual models for English-Spanish and English-German, released in March and April 2024 respectively. That version provides a broader LLM collection for the five largest European languages: English, German, French, Spanish, and Italian. Out of these languages, only German and Spanish are official language directions of the WMT24 shared task and, therefore, the respective bilingual models are chosen for this submission.

The models are based on the Mistral-7B, which was pre-trained for English. In addition, bilingual continual pre-training and subsequent instruction tuning for each language were performed. Both models include the dataset Open-Hermes-2B[1], which contains content in English language and code. The German model `occiglot-7b-de-en-instruct` was trained on 180M tokens of additional multilingual and code instructions, including the German subsets of DiscoLM (which includes the publicly available `germanrag` dataset), Open Assistant Conversations Dataset v2 (OASST-2; Köpf et al., 2023) and Aya-Dataset (Singh et al., 2024). The Spanish model `occiglot-7b-es-en-instruct` was trained on 160M tokens of additional multilingual and code instructions, including the datasets Mentor-ES, the Stanford Question Answering Dataset v2[2] (SQuAD; Carrino et al., 2020) and the Spanish subsets of OASST-2 and Aya-Dataset.

The full instruction fine-tuning took place on an H100 with 8 GPUs for 0.6–4 training epochs (depending on dataset sampling). We used the `axolotl` framework, maintaining a precision of bf16, a global batch size: 128 (with 8192 context length and Cosine Annealing with Warm-up). The tokenizer is unchanged from Mistral-7B-v0.1.

All pre-trained and instruction-tuned checkpoints are available on Hugging Face[3] under the Apache 2.0 license. Note that the model was not safety-aligned and might generate problematic outputs.

[1] https://huggingface.co/teknium

[2] https://huggingface.co/datasets/ccasimiro/squad_es

[3] https://huggingface.co/collections/occiglot/occiglot-eu5-7b-v01-65dbed502a6348b052695e01

3.2 Translation data during training

Both the bilingual German and Spanish models were subjected to paired English translation data during continual pre-training. Specifically, the training data contains paired sentences from Tatoeba (Tiedemann, 2020) and Opus 100 (Zhang et al., 2020). The samples are presented as one coherent text using a diverse set of templates, like

```
Given the following passage:
<German sentence>
a good English translation is:
<English sentence>
```

About 470k and 380k similar translation examples were included during the continual pre-training of the bilingual German and Spanish model, respectively.

Additionally, the instruction tuning stage of both models also includes multilingual data. For the bilingual Spanish model, as mentioned above, parts of the instruction training set were taken from a translated version of the SQuAD, which contains Spanish questions about English literature, for example. More importantly for our task, the incorporated open-assistant OASST-2 dataset also includes about 100 samples of direct instructions for translations between English and Spanish. Similarly, the employed German instruction tuning dataset contains over 2000 dedicated translation examples.

3.3 Prompting translations

During the development of the model, we devised a system prompt instructing the model to perform as a dedicated translator and we found that this prompt is immensely helpful when employing the downstream model for translation tasks. Nevertheless, for the WMT submission we decided to use a prompting method which is similar to the way other LLMs are prompted, so that the results are comparable. Prompting was based on the 5-shot templates used by the organizers General Shared task of Machine Translation to prompt GPT-4[4]. The exact prompt used can be seen in Figure 1.

The suggested practice for MT prompting is multi-shot, where one provides first 4 source/translation samples and then only a source awaiting the translation. Occiglot was giving as an answer not only the translation, but was proceeding with generating more text, on the similar

[4] https://github.com/wmt-conference/wmt23-news-systems/tree/master/tools/LLM-prompt

```
SYSTEM_PROMPT = "You are a very good
translator. Please translate the given
texts from English to 1. target_lang
as precisely and accurately as possible
without changing the structure and answer
only with one translation."

PROMPT = "Please translate this into 1.
{target_lang}:

{source_seg}
1. {translation}"
```

Figure 1: Prompt used

pattern, which was difficult to post-process. We had to write a post-processing script that isolates the translation from the additional superfluous text. Nevertheless, we suspect that this post-processing script may have not operated properly in all cases, as we have some hundreds of empty outputs.

The second issue we faced was the inference speed. We loaded the model locally on a python script in the GPU cluster and used the huggingface `pipeline` command to prompt. The German model was too slow (2-7sec per segment), which made it very tight to meet the deadline. We therefore enabled multiple workers with batches (batch_size=64, num_workers=4) which gave indeed a big acceleration. The behavior of the model was a bit different in the batch mode, so we had to include a system prompt (which was not used for the Spanish model). The parameters of the request command with batches were also different (e.g. the limit max_new_tokens), so it is not sure if parallelizing gave the same results as the single worker mode would have given. The Spanish model was fast enough, and the Spanish test set significantly smaller, so we didn't have to parallelize.

Finally, the German model was going through memory spikes and was killed several times by the administrator rules of our GPU cluster. This may have to do with the test set, as the German test set contains a higher number of examples with more complex sequences. In the future, we have to modify our scripts to stream directly to a file and have the possibility to resume from a particular line in case of a crash.

System Name	AutoRank↓	MetricX↓	Comet Kiwi↑
Unbabel	1.0	1.1	0.723
Dubformer	1.8	1.2	0.694
...			
GPT-4	1.8	1.4	0.700
...			
Mistral-Large	2.0	1.5	0.694
...			
IKUN-C	3.8	2.0	0.641
...			
CUNI-NL	4.2	2.1	0.624
AIST-AIRC	7.2	3.3	0.551
NVIDIA-NeMo †	7.4	3.5	0.558
Occiglot	8.2	3.8	0.539
MSLC	11.9	4.4	0.390
TSU-HITs	13.3	5.6	0.395

Table 1: Indicative comparisons from the preliminary WMT24 General MT automatic ranking for English-German.

System Name	Comet Kiwi ↑
Occiglot	0.539
Mistral 7B v0.1	0.429

Table 2: Comparison between Occiglot and its pre-trained model Mistral7B on English-German

4 Evaluation

4.1 Comparison with other WMT systems

The preliminary results (Kocmi et al., 2024b) of the General MT task, based on automatic measures Table 1, indicate a low performance of Occiglot as compared to other systems. We attribute these results to the fact that the development of our LLM is in the early stage and the model has undergone a relatively minimal optimization for translation. Additionally, we have strong indications that the post-processing script did not account for all possible cases. The fact that the model delivered some hundreds of empty outputs is also a matter that may have contributed to the low scores (although it needs to be noted that the parent model Mistral-Large, prompted by the WMT24 organizers, has delivered a higher number of empty outputs). Finally, we should note that the comparison is mostly done with LLMs with a higher number of parameters, as compared to our system. Therefore, this comparison should only be seen with a grain of salt.

4.2 Comparison with pre-trained model

Occiglot performs better in translating from English-German than the pre-trained model Mistral 7B v0.1, it has been based on. This indicates a

category	items	acc
Ambiguity	22	86.4
Coordination & ellipsis	124	60.5
False friends	40	92.5
Function word	40	75.0
LDD & interrogatives	207	76.3
Lexical Morphology	39	61.5
MWE	123	76.4
Named entity & terminology	112	77.7
Negation	18	66.7
Non-verbal agreement	109	87.2
Punctuation	37	51.4
Subordination	191	85.3
Verb semantics	23	60.9
Verb tense/aspect/mood	3249	71.9
Verb valency	114	65.8
micro-average	4448	72.8
macro-average	4448	73.0

Table 3: Performance of the Occiglot English-German model with regard to linguistically-motivated categories

success of the bilingual continual pre-training and subse- quent instruction tuning for this particular language direction.

4.3 Fine-grained linguistic analysis

Additionally to the automatic scores, we provide here some fine-grained analysis based on particular linguistic categories, based on a linguistically-motivated test suite (Macketanz et al., 2022, 2021; Avramidis et al., 2020). The results can be seen in Table 3 and a more detailed view of the phenomena is displayed in Table 4. The model is particularly strong in *false friends*, which typically refers to lexemes that are identical in their phonological or orthographic form across two languages but have different meanings. It also performs relatively well in handling *non-verbal agreement*, i.e. ensuring that nouns and pronouns agree in gender, number and sometimes case across the sentence (particularly *substitution and coreference*), as well as in *lexical ambiguity*, where a word changes its meaning depending on a context, and *subordination* (particularly *adverbial and subject clause*). *Subordination* refers to the relationship between clauses where one clause is syntactically dependent on the main clause. However, it performs poorly in *punctuation* and particularly quotation marks, which means the model fails to correctly mark direct speech, quotations, or special terms. The low accuracy in *negation* is also particularly concerning, given the semantic importance of this category.

5 Conclusion and further work

We presented an entry participation of a new open-source community-based LLM. Despite some efforts to improve our LLM performance towards translation, the resulting model performs poorly as compared to other systems. Nevertheless, the challenges served as a motivation to unite community forces and initiate research on a new LLM task, which may be further improved in the future. Aside from the automatic scores, by applying a linguistically motivated test suite, we could gain some insights into the linguistic categories which perform better or worse. Further work may include more optimization towards translation, improvement of the prompting and post-processing mechanism and addition of more languages. A more direct comparison with models of similar parameter size (7B) should also be considered in the future.

Acknowledgements

We would like to thank Disco Research, Jan Philipp Harries, and Björn Plüster for making their dataset available to us. Model training was supported by a compute grant at the 42 supercomputer, a central component in the development of hessian AI, the AI Innovation Lab, funded by the Hessian Ministry of Higher Education, Research and the Art (HMWK) & the Hessian Ministry of the Interior, for Security and Homeland Security (HMinD), and the AI Service Centers, funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK). The curation of the training data is partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project OpenGPT-X (project no. 68GX21007D). The linguistic evaluation has been supported by the German Research Foundation (DFG) through the project TextQ.

phenomenon	items	acc
Ambiguity	22	86.4
Lexical ambiguity	22	86.4
Coordination & ellipsis	124	60.5
Gapping	20	25.0
Pseudogapping	19	73.7
Right node raising	18	88.9
Sluicing	20	75.0
Stripping	23	39.1
VP-ellipsis	24	66.7
False friends	40	92.5
Function word	40	75.0
Focus particle	23	78.3
Question tag	17	70.6

phenomenon	items	acc
LDD & interrogatives	207	76.3
Extraposition	18	55.6
Inversion	27	77.8
Multiple connectors	20	80.0
Negative inversion	20	80.0
Pied-piping	19	73.7
Polar question	18	77.8
Preposition stranding	19	57.9
Split infinitive	19	94.7
Topicalization	20	80.0
Wh-movement	27	81.5
Lexical Morphology	39	61.5
Functional shift	17	70.6
Noun formation (er)	22	54.5
MWE	123	76.4
Collocation	20	90.0
Compound	16	87.5
Idiom	20	40.0
Nominal MWE	20	75.0
Prepositional MWE	18	83.3
Verbal MWE	29	82.8
Named entity & terminology	112	77.7
Date	19	73.7
Domainspecific Term	18	83.3
Location	19	84.2
Measuring unit	21	76.2
Onomatopeia	15	53.3
Proper name	20	90.0
Negation	18	66.7
Non-verbal agreement	109	87.2
Coreference	35	88.6
Genitive	18	83.3
Personal Pronoun Coreference	13	92.3
Possession	27	81.5
Substitution	16	93.8
Punctuation	37	51.4
Quotation marks	37	51.4
Subordination	191	85.3
Adverbial clause	19	94.7
Cleft sentence	17	76.5
Contact clause	22	72.7
Indirect speech	19	89.5
Infinitive clause	19	84.2
Object clause	20	95.0
Pseudo-cleft sentence	19	78.9
Relative clause	39	89.7
Subject clause	17	82.4
Verb semantics	23	60.9
Verb tense/aspect/mood	3249	71.9
Conditional	20	70.0
Ditransitive - conditional I progressive	53	71.7
Ditransitive - conditional I simple	55	76.4
Ditransitive - conditional II progressive	56	48.2
Ditransitive - conditional II simple	54	77.8
Ditransitive - future I progressive	52	86.5
Ditransitive - future I simple	110	70.0
Ditransitive - future II progressive	55	34.5
Ditransitive - future II simple	51	29.4
Ditransitive - past perfect progressive	56	62.5
Ditransitive - past perfect simple	55	67.3
Ditransitive - past progressive	57	77.2
Ditransitive - present perfect progressive	57	75.4
Ditransitive - present perfect simple	51	80.4
Ditransitive - present progressive	55	85.5
Ditransitive - simple past	76	85.5
Ditransitive - simple present	50	84.0
Gerund	25	80.0
Imperative	15	46.7
Intransitive - conditional I progressive	27	92.6
Intransitive - conditional I simple	28	96.4
Intransitive - conditional II progressive	27	66.7
Intransitive - conditional II simple	29	69.0
Intransitive - future I progressive	30	83.3
Intransitive - future I simple	68	91.2
Intransitive - future II progressive	28	53.6
Intransitive - future II simple	35	48.6
Intransitive - past perfect progressive	30	46.7
Intransitive - past perfect simple	35	71.4
Intransitive - past progressive	32	81.3
Intransitive - present perfect progressive	29	82.8
Intransitive - present perfect simple	29	72.4
Intransitive - present progressive	61	85.2
Intransitive - simple past	35	80.0
Intransitive - simple present	38	68.4
Modal	288	71.5
Modal negated	304	75.0
Reflexive - conditional I progressive	35	74.3
Reflexive - conditional I simple	34	64.7
Reflexive - conditional II progressive	34	58.8
Reflexive - conditional II simple	34	76.5
Reflexive - future I progressive	30	60.0
Reflexive - future I simple	68	54.4
Reflexive - future II progressive	34	41.2
Reflexive - future II simple	33	39.4
Reflexive - past perfect progressive	35	42.9
Reflexive - past perfect simple	34	67.6
Reflexive - past progressive	33	87.9
Reflexive - present perfect progressive	32	68.8
Reflexive - present perfect simple	34	79.4
Reflexive - present progressive	33	75.8
Reflexive - simple past	33	78.8
Reflexive - simple present	31	61.3
Transitive - future II progressive	30	36.7
Transitive - conditional I progressive	30	86.7
Transitive - conditional I simple	27	85.2
Transitive - conditional II progressive	28	89.3
Transitive - conditional II simple	25	80.0
Transitive - future I progressive	30	73.3
Transitive - future I simple	57	84.2
Transitive - future II simple	32	65.6
Transitive - past perfect progressive	28	89.3
Transitive - past perfect simple	28	71.4
Transitive - past progressive	44	70.5
Transitive - present perfect progressive	27	88.9
Transitive - present perfect simple	29	79.3
Transitive - present progressive	39	84.6
Transitive - simple past	38	89.5
Transitive - simple present	34	88.2
Verb valency	114	65.8
Case government	14	85.7
Catenative verb	18	83.3
Mediopassive voice	22	54.5
Passive voice	19	78.9
Resultative	19	63.2
Semantic roles	22	40.9
micro-average	4448	72.8
phen. macro-average	4448	73.2
categ. macro-average	4448	73.0

Table 4: Performance of the Occiglot English-German model with regard to linguistically-motivated phenomena

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context Examples Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish Translation of SQuAD Dataset for Multi-lingual Question Answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Ryandito Diandaru, Lucky Susanto, Zilu Tang, Ayu Purwarianti, and Derry Wijaya. 2024. Could We Have Had Better Multilingual LLMs If English Was Not the Central Language? ArXiv:2402.13917 [cs].

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. ArXiv:2404.13813 [cs].

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. 2024a. 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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024b. Preliminary wmt24 ranking of general mt systems and llms.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Advances in neural information processing systems*, volume 36, pages 47669–47681. Curran Associates, Inc.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 936–947, Marseille, France. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Abdelhadi Soudi, Mohamed Hannani, Kristof Van Laerhoven, and Eleftherios Avramidis. 2024. Exploring the potential of large language models in adaptive machine translation for generic text and subtitles. In *Proceedings of the 17th workshop on building and using comparable corpora (BUCC) @ LREC-COLING 2024*, pages 51–58, Torino, Italia. ELRA and ICCL.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. ArXiv:2405.20089 [cs].

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference *on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Changtong Zan, Liang Ding, Li Shen, Yibing Zhen, Weifeng Liu, and Dacheng Tao. 2024. Building Accurate Translation-Tailored LLMs with Language Aware Instruction Tuning. ArXiv:2403.14399 [cs].

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41092–41110. PMLR. ISSN: 2640-3498.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.