# Conditioning LLMs with Emotion in Neural Machine Translation

**Charles Brazier** and **Jean-Luc Rouas**
Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
charles.brazier@u-bordeaux.fr jean-luc.rouas@labri.fr

## Abstract

Large Language Models (LLMs) have shown remarkable performance in Natural Language Processing tasks, including Machine Translation (MT). In this work, we propose a novel MT pipeline that integrates emotion information extracted from a Speech Emotion Recognition (SER) model into LLMs to enhance translation quality. We first fine-tune five existing LLMs on the Libri-trans dataset and select the most performant model. Subsequently, we augment LLM prompts with different dimensional emotions and train the selected LLM under these different configurations. Our experiments reveal that integrating emotion information, especially arousal, into LLM prompts leads to notable improvements in translation quality.

## 1 Introduction

Large Language Models (LLMs) are transformer-based (Vaswani et al., 2017) deep learning models designed to understand and generate natural language text by predicting the probability of the next token in a sequence. LLMs excel across various Natural Language Processing (NLP) tasks, such as information retrieval (Zhu et al., 2023b), instruction following (Ouyang et al., 2022), or engaging in chatbot discussions (OpenAI, 2022).

Among NLP tasks, LLMs have shown great capacities in Machine Translation (MT) (Zhu et al., 2023a), the task of translating a text from one language to another. Previous research has enhanced LLM performance in MT through various strategies, including optimized prompting techniques (Zhang et al., 2023), in-context learning features (Brown et al., 2020) to improve translation quality over time (Moslem et al., 2023a,b), and a two-stage fine-tuning method composed of a first fine-tuning on monolingual data to learn general linguistic knowledge followed by a second fine-tuning on parallel data (Xu et al., 2023) that establishes the current state-of-the-art method in MT.

Apart from LLMs, previous works in MT have demonstrated the possibility of controlling the translation by adding extra information to the model that is not explicitly specified in the source sentence to be translated, and that can influence the translation. Existing works in that direction focused on the control of politeness (Sennrich et al., 2019), gender (Vanmassenhove et al., 2018; Gaido et al., 2023), or emotion (Brazier and Rouas, 2024) of the translation and showed that this extra information helps improve translation quality.

In this work, we propose to improve translation performances of an LLM-based model by adding emotion as extra information in the prompt of the model to condition the translation. This work relies on the fact that words can be classified into emotion categories, leading to affective word lists (Pennebaker et al., 2001). Thus, conditioning the translation with a specific emotion would use a suitable vocabulary in the translation. In Brazier and Rouas (2024), authors showed that adding arousal information, reflecting the level of stimulation (ranging from calm to excited), extracted from the voice and added at the start of each input text sentence, helps improve translation performances. In the following, we study the behavior of several LLMs for the task of MT when emotion dimensions are added to input prompts.

To address this problem, we first fine-tune several existing LLMs for the task of English-to-French text-to-text translation. Then, after selecting the best model as baseline for our experiments, we compute for each input sentence its emotional dimensions with the help of a state-of-the-art Speech Emotion Recognition (SER) model applied to audio recordings. Finally, we compare translation performance with and without the addition of each emotional dimension as extra information added to each input prompt. We show that emotion improves translation (BLEU and COMET), especially in the case of arousal.

*Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 81–86
August 15-16, 2024 ©2024 Association for Computational Linguistics

## 2 Related works

In this work, we aim at combining an LLM-based MT model with emotion information to improve translation performances. In the following, we first describe a close work that performs this combination without the use of an LLM. Then, we list several existing LLMs that can be used as a baseline for our MT task.

### 2.1 Machine Translation with Emotion

To our knowledge, the only work that combines an MT model with emotion information is described in Brazier and Rouas (2024). In this study, the authors utilize a state-of-the-art Speech Emotion Recognition (SER) model (Wagner et al., 2023) to automatically estimate dimensional emotion values, including arousal, dominance, and valence, for each audio recording associated with text sentence. These values are then transformed into unique emotion tokens, either positive or negative, which are added at the beginning of tokenized input text sentences. The authors report an increase in translation BLEU score, especially when adding arousal tokens at the start of input sentences.

The MT model used for their experiments is a transformer-based encoder-decoder architecture, comprising 6 layers for the encoder, 6 layers for the decoder, and 4 attention heads in each self-attention layer. The model is trained on the Libri-trans dataset (Kocabiyikoglu et al., 2018), which includes triplets of English recordings, English texts, and French texts, totaling 235 hours of data (230h for train, 2h for dev, and 3.5h for test). The model performs English-to-French translation.

In this work, we propose to use the same translation pipeline, but instead of using a specific MT model, we replace it with a fine-tuned LLM. Since LLMs have more trainable parameters, we anticipate improved translation performances. However, our objective is to observe how LLMs behave when augmented with emotion information in the input prompt.

### 2.2 LLM selection for MT

Recent advances in Large Language Modeling have significantly expanded the capabilities of LLMs across various tasks, such as reasoning, coding, or mathematics. Among the numerous existing LLMs (Chiang et al., 2024), the best-performing models are GPT-4 (OpenAI, 2023), LLaMA 3 (AI@Meta, 2024), Gemini 1.5 (Team, 2024), or Claude 3 (Anthropic, 2024).

For the task of MT, we restrict our LLM selection to models that are open-source, promising (high rank in the LLM arena[1], or already fine-tuned to the MT task), and that only contain 7 billion (7B) of parameters. We select 5 different models that are described in the following.

The first selected LLM is *Mistral-7B-v0.1*[2], an open-source model (Jiang et al., 2023) which ranks among the best 7B-parameter models.

As the second model, we select *Mistral-7B-Instruct-v0.2*[3]. The model is similar to the previous model but has been fine-tuned to follow instructions.

Our third selected model is *TowerBase-7B-v0.1*[4]. This model (Alves et al., 2024) is based on LLaMA 2 (AI@Meta, 2023) and its training has been continued on multilingual data (including English and French monolingual data, as well as bilingual data).

Similarly to Mistral, we select *TowerInstruct-7B-v0.2*[5] as our fourth model. This model is a variant of the previous one that has been fine-tuned to follow instructions including translations.

Finally, as our fifth model, we select the SOTA MT model *ALMA-7B-R*[6], which is based on LLaMA 2 (AI@Meta, 2023), and fine-tuned on monolingual and parallel data. However, the data used for fine-tuning does not include French.

## 3 Experiments and results

In this section, we describe our experiments for the task of English-to-French text-to-text translation. We conduct two successive experiments. Firstly, we fine-tune five existing LLMs on the Libri-trans dataset (Kocabiyikoglu et al., 2018) and consider the best model as a foundation for our second experiment. Secondly, we fine-tune the selected LLM on the same task but under different configurations. Henceforth, prompts used for translation include each emotion dimension that is automatically estimated from the SER model.

### 3.1 Fine-tuning LLMs on Libri-trans

To perform MT with LLMs, the task needs to be converted into a language modeling problem with

[1] http://chat.lmsys.org/?leaderboard

[2] http://huggingface.co/mistralai/Mistral-7B-v0.1

[3] http://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[4] http://huggingface.co/Unbabel/TowerBase-7B-v0.1

[5] http://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

[6] http://huggingface.co/haoranxu/ALMA-7B-R

| Model | BLEU | | COMET | |
|---|---|---|---|---|
| | dev | test | dev | test |
| Mistral | 16.4 | 16.7 | 73.2 | 72.5 |
| MistralInstruct | 16.0 | 17.9 | 72.1 | 71.9 |
| TowerBase | **24.0** | **20.6** | **73.8** | **72.9** |
| TowerInstruct | 6.4 | 6.1 | 35.5 | 35.5 |
| ALMA | 7.1 | 7.5 | 52.1 | 52.8 |

Table 1: BLEU and COMET scores of our five selected LLMs on dev and test sets of Libri-trans.

the use of prompts. In this work, we perform zero-shot prompting and follow two different templates. The first template will be applied to *Mistral-7B-v0.1* and *TowerBase-7B-v0.1*:

```
English: <src txt> \n French: <tgt txt>
```
(1)

where `<src txt>` and `<tgt txt>` refer to the English source sentence and the French target sentence respectively.

The second template will be applied to models that follow instructions, namely *Mistral-7B-Instruct-v0.2*, *TowerInstruct-7B-v0.2*, and *ALMA-7B-R*:

```
[INST] Translate from English to French: <src txt> [/INST] \n <tgt txt>
```
(2)

To fine-tune LLMs, we employ QLoRA (Hu et al., 2022; Dettmers et al., 2023), a Parameter Efficient Fine-Tuning method (Mangrulkar et al., 2022) that allows training with significantly fewer parameters. Additionally, we apply a 4-bit quantization to reduce memory usage while maintaining 16-bit precision during computation.

We provide two distinct metrics to evaluate our MT models. The first metric is the BLEU score computed using sacrebleu (Post, 2018). It reflects the degree of lexical matches (number of common n-grams) between the proposed translation and its corresponding reference. The second metric is the COMET score [7] (Rei et al., 2022). It is computed from a trained model and reflects translation quality between translation, reference, and also the source sentence. According to the metric ranking presented in Freitag et al. (2022), we rely more on the COMET score than on the BLEU score.

Table 1 showcases the results of our first experiment. In this table, we report BLEU and COMET scores of the five selected LLMs on both the dev and test sets of the Libri-trans dataset.

The table highlights three models, *Mistral-7B-v0.1*, *Mistral-7B-Instruct-v0.2*, and *TowerBase-7B-v0.1*, that attain high BLEU and COMET scores. They obtain COMET scores ranging from 72.1 to 73.8 on the dev set and from 71.9 to 72.9 on the test set. Additionally, their BLEU scores ranged from 16.0 to 24.0 on the dev set and from 16.7 to 20.6 on the test set. While COMET scores are not meant to be interpretable (but enable the comparison between models), BLEU scores indicate, on average, a translation that is more or less clear with numerous grammatical errors. These low BLEU scores are comparable to performances of previous works on this dataset (Zhao et al., 2021; Brazier and Rouas, 2024) and are mainly caused by the nature of the data (audiobooks with literary vocabulary).

Also, it is worth noting that two models, *TowerInstruct-7B-v0.2* and *ALMA-7B-R*, exhibit poor performances in MT when fine-tuned on Libri-trans. In the case of *ALMA-7B-R*, this can be explained by the fact that French is not among the languages included in the data used to pre-train the model. Thus, the model fails at predicting French text.

As additional training information, all LLMs have obtained their optimal state in a maximum of 5 epochs. This represents a training time of 3 hours on a GPU NVIDIA A100 for each model. This fast fine-tuning time is due to QLoRA and 4-bit quantization strategies.

To summarize, the best machine translation performances were achieved with the *TowerBase-7B-v0.1*. This LLM serves as a baseline and foundation model for the following experiment.

### 3.2 Fine-tuning LLMs with Emotion

The second experiment aims at observing the behavior of our LLM-based *TowerBase-7B-v0.1* model on the task of English-to-French Machine Translation when emotion information is added to the prompt before translation.

As a first step, we estimate the emotion of each English recording present in the Libri-trans dataset. Following the same methodology as Brazier and Rouas (2024), we compute dimensional emotion values for arousal, dominance, and valence with the help of a trained SER model (Wagner et al., 2023). Emotion values range between 0 and 1 and are correctly balanced (medians between 0.4 and 0.6, see Brazier and Rouas (2024)).

As a second step, we create specific prompts that include the emotion information in the text. For

[7] https://huggingface.co/Unbabel/wmt22-comet-da

this purpose, we propose 3 different templates. The first template adds emotion information before the source sentence:

$$\texttt{English <status> <emotion>: <src txt> \textbackslash n French: <tgt txt>} \quad (3)$$

where `status` is replaced by either *with* or *without* if the emotion value is higher or lower than 0.5 respectively, `emotion` is replaced by either *arousal*, *dominance*, or *valence*, `src txt` represents the English source sentence, and `tgt txt` represents the French target translation.

The second template adds emotion information before the target sentence:

$$\texttt{English: <src txt> \textbackslash n French <status> <emotion>: <tgt txt>} \quad (4)$$

The third template is inspired from Brazier and Rouas (2024), where emotion information is added as a discrete token at the start of the source sentence:

$$\texttt{English: [<emotion> <polarity>] <src txt> \textbackslash n French: <tgt txt>} \quad (5)$$

where `polarity` is replaced by either *positive* or *negative* if the emotion value is higher or lower than 0.5 respectively.

In this experiment, the *TowerBase-7B-v0.1* model is retrained from its initial state and not from the training checkpoint obtained after the previous experiment. In the following, all models obtain their best performances in less than 5 training epochs.

Table 2 showcases the results of our second experiment. It reports BLEU and COMET scores of the selected *TowerBase-7B-v0.1* model on the dev and test sets of the Libri-trans dataset under different configurations. The first line mentions the score of the LLM obtained in the previous experiment and serves as a baseline for the second experiment. The other lines correspond to the model trained with different emotions (arousal, dominance, or valence), and with different prompts (the numbers 3, 4, and 5 refer to their equation number).

We first remark that, except in the case of *dominance5*, all COMET scores improved, compared to their baseline. This reflects a better translation quality when adding emotion information to the prompts. The best COMET scores are obtained when arousal information is added to the prompt using Equation 3. In this configuration, COMET scores are increased by +1.1 and +1.4 for the dev and test sets of Libri-trans respectively.

| **Model** | **BLEU** | | **COMET** | |
|---|---|---|---|---|
| | **dev** | **test** | **dev** | **test** |
| TowerBase | 24.0 | 20.6 | 73.8 | 72.9 |
| +arousal3 | 22.1 | 21.8 | **74.9** | **74.3** |
| +arousal4 | **25.6** | **24.1** | 74.8 | 73.9 |
| +arousal5 | 19.3 | 19.2 | 74.2 | 73.4 |
| +dominance3 | 19.9 | 19.4 | 74.4 | 73.5 |
| +dominance4 | 18.9 | 20.9 | **74.9** | 74.0 |
| +dominance5 | 16.5 | 20.1 | 73.4 | 73.0 |
| +valence3 | 21.5 | 18.9 | 74.1 | 73.5 |
| +valence4 | 18.3 | 21.2 | 74.6 | 73.9 |
| +valence5 | 17.2 | 16.0 | 74.5 | 73.6 |

Table 2: BLEU and COMET scores of the TowerBase model on dev and test sets of Libri-trans. First line: baseline score. Other lines: score when trained with emotion in the prompt.

Secondly, we observe that BLEU scores show improvements only for specific models. The best BLEU scores are obtained when arousal information is added to the prompt using Equation 4. In this configuration, BLEU scores increase by +1.6 and +3.5 for the dev and test sets of Libri-trans respectively. However, due to the low ranking of BLEU (Freitag et al., 2022), we do not conduct further analysis based on this metric.

In summary, incorporating emotion information into the translation process appears to enhance translation quality. The highest scores are achieved when utilizing the arousal dimension with Equation 3 or 4. This finding aligns with the results reported in Brazier and Rouas (2024).

## 4 Conclusion

We proposed a new MT pipeline that combines an LLM-based model and emotion information extracted from a SER model to improve translation performances. We obtain the best performances when the arousal value is added to the LLM prompt.

As future work, we will apply our method to other multilingual datasets including Must-C (Di Gangi et al., 2019). Unlike the Libri-trans dataset, which consists of literary text read by speakers, Must-C encompasses various speech types, such as TED talks, which can offer more emotional variability and therefore further enhance translation performance. We also plan to extend our method to the speech-to-text task, also known as Speech translation.

## 5 Acknowledgements

The research presented in this paper is conducted as part of the project FVLLMONTI, which has received funding from the European Union's Horizon 2020 Research and Innovation action under grant agreement No 101016776.

## References

AI@Meta. 2023. LLaMA 2: Open Foundation and Fine-tuned Chat Models. *Preprint*, arXiv:2307.09288.

AI@Meta. 2024. LLaMA 3 Model Card.

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.

Anthropic. 2024. Claude 3: Introducing the Next Generation of Claude.

Charles Brazier and Jean-Luc Rouas. 2024. Usefulness of Emotional Prosody in Neural Machine Translation. In *Proc. of the International Conference on Speech Prosody (SP)*, Leiden, The Netherlands.

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, Daniel M. Ramesh, Aditya ans Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Bernet, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-shot Learners. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, Virtual.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *Preprint*, arXiv:2403.04132.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient Finetuning of Quantized LLMs. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, New Orleans, LA, USA.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2012–2017, Minneapolis, MN, USA.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proc. of the Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates.

Marco Gaido, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. How to Build Competitive Multi-gender Speech Translation Models for Controlling Speaker Gender Translation. In *Proc. of the Italian Conference on Computational Linguistics (CLiC-it)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of the International Conference on Learning Representations (ICLR)*, Virtual.

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

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. https://github.com/huggingface/peft.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive Machine Translation with Large Language Models. In *Proc. of the Annual Conference of the European Association for Machine Translation (EAMT)*, pages 227–237, Tampere, Finland.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023b. Fine-tuning Large Language Models for Adaptive Machine Translation. *Preprint*, arXiv:2312.12740.

OpenAI. 2022. https://chat.openai.com/.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744, New Orleans, LA, USA.

J.W. Pennebaker, M.E. Francis, and R.J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proc. of the Conference on Machine Translation: Research Papers (WMT)*, pages 186–191, Brussels, Belgium.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proc. of the Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2019. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 35–40, San Diego, USA.

Gemini Team. 2024. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. *Preprint*, arXiv:2403.05530.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (CEMNLP)*, pages 3003–3008, Brussels, Belgium.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 5998–6008, Long Beach, USA.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10745–10759.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. *Preprint*, arXiv:2309.11674.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 41092–41110, Edinburgh, Scotland.

Chengqi Zhao, Mingxuan Wang, Qianqian Dong, Rong Ye, and Lei Li. 2021. NeurST: Neural speech translation toolkit. In *Proc. of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL)*, pages 55–62, Online.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023a. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *Preprint*, arXiv:2304.04675.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023b. Large Language Models for Information Retrieval: A Survey. *Preprint*, arXiv:2308.07107.