Beyond Parallel Corpora: Assessing LLMs and State-of-the-Art Models for Specialised Texts Translation and Error Detection

Anonymous ACL submission

Abstract

MT of specialised texts poses particular challenges due to domain-specific terminology, phraseology and structural conventions. LLMs offer a promising alternative to traditional MT approaches, especially in domains with limited parallel corpora (specialised texts being a notable example). However, their performance remains underexplored, despite the fact that this type of translation has significant socio-economic implications. In this study, we evaluate the ability of LLMs and state-of-the-art translation models to translate specialised texts, using an error typology designed for the evaluation of specialised translation to provide a qualitative assessment. Our approach provides detailed insights into translation challenges and investigates whether LLMs can also detect errors in LSP translations.

1 Introduction

Specialised translation refers to the translation of technical, scientific or professional texts written in specific language often referred to as LSP (Language for Specific Purposes). They are characterised by precise terminology and phraseology, specific textual conventions and specific discourse structures. This type of translation plays a crucial economic role: In a globalised world, the dissemination of scientific knowledge, international expert communication and the development of specialised markets depend heavily on high-quality translations. However, specialised translation is a particularly demanding task. It requires not only a deep understanding of the subject matter, but also the ability to convey concepts that may not exist in the target language, sometimes requiring the creation of new terms, as well as cross-cultural and cross-linguistic text type knowledge (Rogers, 2015).

Traditional MT systems, such as phrase-based and neural MT systems, which rely primarily on supervised learning approaches, face a major obstacle in this context: the scarcity of parallel corpora in specialised domains (Bouamor and Sajjad, 2018). Unlike general language texts, specialised corpora are often limited in size, heterogeneous and expensive to produce (Aston, 1999). LLMs, which are trained on large monolingual and multilingual corpora, offer a promising alternative. Their ability to generalise from unaligned data and reason about linguistic structures allows them to approach translation in a fundamentally different way. Moreover, their creative flexibility could be valuable in adapting or inventing terms that do not exist in the target language.

In this study, we propose to evaluate the ability of LLMs to evaluate specialised translated texts: Our first contribution consists of an evaluation of MT on a corpus of scientific abstracts in NLP. Conducting this type of evaluation presents two main challenges: (i) the lack of an annotated translation corpus in this domain, and (ii) the uncertainty regarding the reliability of automatic metrics for these specific text types. To address these issues, we propose an evaluation approach that prioritises manual evaluation over automatic scoring. Instead of assigning a numerical score to translations, we adopt an error typology and ask an expert to identify and categorise errors. This approach not only overcomes the limitations of automated metrics, but also provides a more interpretable and qualitatively rich assessment of translation quality. Our results (Section 3) show that existing general-public translation systems are still a long way from being able to translate this type of text: the particularities of LSPs remain challenging, leading to translations of insufficient quality.

This methodological choice also led us to explore another aspect of LLMs' abilities to understand and manipulate LSP texts: their ability to detect errors in specialised translations. Using prompt-based evaluation, we assess whether different models can reproduce the annotations made by our expert, pro-

viding new insights into their understanding of specialised language phenomena. Experiments, reported in Section 4, show that at least `ChatGPT-4o` is able to identify a good proportion of LSP translation errors, suggesting it is trained on sufficient specialised data to provide some help to translators in their translation or quality assessment tasks.

2 A Corpus for Identifying Translation Errors in LSPs

An Error Typology for LSP Translation There are two main approaches to evaluating MT quality. The first consists in assigning numerical scores to translations based on overall quality assessments or comparisons with reference corrections. The second approach involves identifying specific errors in the translated texts and categorising them using labels that describe the nature of each error. In this study, we adopt the latter method, as it allows for a qualitative analysis of translation errors. This approach is particularly beneficial for determining whether errors arise from the inherent limitations of the MT system or from the linguistic and domain-specific characteristics of the texts we are analysing, specifically in the context of LSP.

To systematically annotate translation errors, we use a translation error typology. In this context, "typology" refers to a structured classification system that organises translation errors into a hierarchical framework with varying levels of granularity, depending on the degree of precision of the error types contained in the typology. At the highest level, our typology consists of three primary error categories:

- Content Transfer errors, which encompass 6 subcategories and a total of 9 individual error types;
- Language errors, divided into 8 subcategories for a total of 28 error labels, 10 of them being terminological errors;
- Tool-related errors, which contain 4 error types (see Figure 2 in Appendix A for the full typology).

This hierarchical organisation allows for a more precise identification of error types and facilitates targeted improvements in MT models and workflows. Our typology is based on MeLLANGE (Castagnoli et al., 2011; Kübler, 2008), a typology designed for pedagogical purposes in translation training. This typology was adapted to evaluate human and machine translation, along with post-editing quality. We then extended it to detect system-induced errors and enriched it with several error types from the Multidimensional Quality Metrics typology (MQM) introduced in Burchardt (2013), a flexible typology developed as part of the QTLaunchPad project to assess the quality of machine and human translation across domains.

Among the 41 error types contained in our typology, 11 errors are directly related to the specialised nature of the texts. These errors can be caused by various aspects, such as the suitability of the register expected in LSPs, the domain-specific terminology and phraseology, as well as terminological inconsistencies. The remaining 30 error types are generally related to the translation process, including overly literal translations or distortions of the source text's meaning. By systematically categorising and analysing these errors, we gain valuable insights into the challenges posed by MT in specialised domains and can propose strategies for improving translation quality in these contexts.

A Corpus of MT of NLP Articles In all our experiments, we considered a corpus specifically created for this study. This corpus consists of 35 abstracts of articles in English published in NLP conference proceedings.[1] On average, these abstracts contain 9 sentences and 159 words, totalling 331 sentences and 5,718 words.[2] While this corpus is smaller than those usually considered in NLP research, it includes detailed manual annotations that can only be performed by an expert and are particularly time-consuming to collect.

These articles were automatically translated into French using two publicly available machine translation systems: `DeepL` and `ChatGPT`. `DeepL` is a commercial translation model available online, while `ChatGPT` is a general-purpose LLM not specifically designed for translation but frequently used for this purpose.[3] `ChatGPT`'s translation capabilities have already been the subject of several publications (cf. e.g. Lyu et al. (2024); Wang et al. (2023); Siu (2023); Jiao et al. (2023); He (2024)).

[1] All these articles are distributed under a CreativeCommons free license. Our corpus will be distributed upon publication.

[2] Segmentation in sentences and tokenization in words have been done using `spacy` with the `en_core_web_sm` model (Honnibal and Montani, 2017).

[3] We used a prompt in French to ask `ChatGPT` to translate the text. Here is the translation of the prompt in English: "You are a translator who specialises in translating research articles on natural language processing. Translate the following text into French, respecting the structure of the original text and not omitting any elements."

Exploring commercial systems with limited publicly available information may seem contrary to a scientific approach. But, our choice is driven by practical considerations: both systems are widely used by professional translators in their work, making it essential to accurately assess their capabilities.

A professional translator, specialised in translation evaluation and experienced in NLP translation and evaluation, annotated all of DeepL translations and 25 (out of 35) translations by ChatGPT following the typology we have just introduced. An example of reference annotation is provided in Figure 1. This example illustrates that annotations involve both identifying errors in the abstract and describing them by assigning one or more labels. While it is challenging to precisely measure the time required for this annotation work, we estimate that annotating one abstract takes between 30 and 90 minutes, depending on the length of the text and the level of difficulty/specialisation.

3 Evaluating State-of-the-Art Translation Models' Ability to Translate LSP

The annotations produced by our expert allow us to assess the overall ability of the two translation systems to translate LSP. A simple preliminary metric shows that the translations produced by ChatGPT contain an average of 1.2 errors per sentence, whereas those produced by DeepL contain 1.8 errors per sentence. Although the translations produced by ChatGPT clearly contain fewer errors, the quality of both systems is far from satisfactory: a translator still has to correct at least one error per sentence, a considerable amount of work.

The labels assigned by the expert provide a more detailed understanding of the nature of the errors. Each error received between 1 and 6 different labels (2.24 ± 0.08 on average,[4] 2.05 ± 0.11 for translation by ChatGPT and 2.31 ± 0.09 for those of DeepL). Of the 41 error labels defined in the error typology considered, 38 were used at least once. The five most frequently used labels are for each MT system considered are reported in Table 1. It is interesting to note that both systems tend to make similar errors, primarily involving terminology. More broadly, 41.5 % of ChatGPT's errors include at least one label related to the specificity of LSPs, while this proportion increases to 51.3 % for DeepL (overall 58.9 % of the errors include at least one error type specific to LSPs). The mere fact that an error annotated with several different labels includes at least one LSP-related error label is sufficient to classify it as an LSP error; in fact, LSP errors are often associated with other labels that describe the impact of the LSP error (for example, a terminology error can distort the meaning of the source text, thus resulting in two labels: terminology + distortion).

[4] In all reported results, we include the 95% CI intervals calculated using using the bca bootstrap method of Efron and Tibshirani (1993).

Error label	% Error with this label
ChatGPT	
Akward Style (LA-ST-AW)	40.4 %
Incorrect terminology (LA-TL-INS) 🛠️	39.9 %
Too litteral (TR-SI-TL)	36.5 %
Incorrect lexis (LA-TL-ING)	16.9 %
Inappropriate collocation (LA-TL-ICS) 🛠️	10.7 %
DeepL	
Incorrect terminology (LA-TL-INS) 🛠️	46.6 %
Akward Style (LA-ST-AW)	17.8 %
Too litteral (TR-SI-TL).	16.9 %
Distortion (TR-DI)	15.7 %
Incorrect lexis (LA-TL-ING)	15.2 %

Table 1: Most frequent error labels. Labels in bold only appear in one of the two lists and 🛠️ indicates an error specific to LSPs. 'Awkward style' is an error marked by unidiomatic word sequences that don't fit lexical, terminological, collocational, or syntactic norms, resulting in unnatural style. 'Incorrect terminology' is the mistranslation of a specialised term. 'Too literal' is a stylistic error when the translation too closely mirrors the source. 'Incorrect lexis' refers to a mistranslated general-language term. 'Inappropriate collocation' is the mistranslation of a specialised collocation. 'Distortion' is a content transfer error altering the source's meaning.

4 Detecting Error in LSP Translations

Context In addition to this first task, we considered a secondary task to assess LLMs' ability to understand and manipulate LSPs: Identifying errors in the translation of technical texts. This task automates the annotations from the previous section. Intuitively, it is simpler as it only requires the detection of 'mismatches' between a source text and its translation, without translating it correctly.

This task has considerable practical value as it can assist translators in their daily work. Furthermore, although we do not specifically explore this aspect in our study, it could also assist translation trainers in correcting and marking students' work. Because of its practical importance, it has recently received increasing attention.

Prompts for Error Detection Following the approach proposed by Fernandes et al. (2023), and

Une analyse contrastive des techniques d'évaluation de la traduction automatique Dans ce chapitre, une enquête LA-TL-INS, LA-TL-ING, LA-TL-FC est présentée TR-SI-TL, LA-ST-AW sur les différentes manières d'évaluer la sortie TR-SI-TL, LA-TL-INS, LA-IA-NU de la traduction automatique. Les méthodologies présentées incluent l'évaluation de la qualité par des évaluateurs humains, les techniques d'évaluation automatisée, l'évaluation sur la base d'une analyse des erreurs LA-TL-INS, LA-TL-ICS et sur la base LA-ST-AW, LA-ST-TA du temps de post-édition, et elles sont mises à l'épreuve LA-ST-AW, LA-TL-ING sur un corpus d'échantillon LA-TL-INS.

Figure 1: Example of human reference annotation, each error being identified by its span (text on an orange background) and one or several labels (in subscript).

recognising the lack of sufficiently large corpora for fine-tuning and other ML methods, we used prompt-based techniques. Specifically, we designed two types of prompts: a simple prompt, which instructed an LLM to identify errors in a zero-shot context (i.e., without providing the model with any information about MT errors or any examples), and a complex prompt, which included our full annotation guidelines and required the model to assign an error category to each detected error. Full prompts can be found in Appendix B.

Evaluation The model's output quality using these prompts was evaluated through precision and recall, measuring predicted errors against gold annotations. Errors were considered matched if they shared at least one character, ensuring the translator's attention to problematic text. We used macro scores, averaging values across documents.

Experimental Results We begin by evaluating the ability of different LLMs to locate errors in the 35 translations of `DeepL`, considering only the simple prompt. The results, reported in Table 2, show that while all the systems achieve relatively high accuracy, their recall remains consistently low. This suggests that they only identify errors they are very confident about. ChatGPT is the only one not to have a low recall, showing once again that this model has a good knowledge of LSPs and the difficulties encountered when translating them.

In order to refine our results, we conducted a second, more detailed experiment in which we used `ChatGPT` not only to locate errors, but also to categorise them according to the typology described in §2. We considered both `DeepL` translations and those produced by `ChatGPT` itself. Our results show that for `DeepL` translations, `ChatGPT` performs these two tasks quite well, achieving a precision of 0.79 ± 0.04 and a recall of 0.65 ± 0.05, with 64.1 % of the labels assigned matching one of those given by our expert. However, the quality of its responses varies considerably across documents

model	P	R	F_1
`LLama2` (7B params. (Touvron et al., 2023))	0.51 ± 0.032	0.21 $\pm$ 0.02	0.29 $\pm$ 0.023
`LLama3` (8B params. (Grattafiori et al., 2024))	0.71 ± 0.074	0.33 $\pm$ 0.044	0.44 $\pm$ 0.054
`Deepseek` (7B params. (DeepSeek-AI et al., 2025))	0.64 ± 0.082	0.25 $\pm$ 0.077	0.32 $\pm$ 0.069
`EuroLLM-instruct` (1.7B params.(Martins et al., 2024))	0.65 ± 0.18	0.19 $\pm$ 0.065	0.29 $\pm$ 0.09
`Mistral` (7B. params. (Jiang et al., 2023))	0.79 ± 0.034	0.2 $\pm$ 0.024	0.29 $\pm$ 0.025
`ChatGPT-4o`	0.75 $\pm$ 0.06	0.67 $\pm$ 0.05	0.70 $\pm$ 0.05

Table 2: Evaluation of the ability of different LLMs to locate translation errors in LSP texts. `EuroLLM` is an LLM fine-tuned on an instruction dataset, with a focus on general instruction-following and machine translation.

(see their distribution in Appendix C), raising concerns about their reliability for professional translators: Inconsistent performance makes post-editing unpredictable, undermining efficiency gains. Furthermore, the results are significantly worse when `ChatGPT` evaluates its own translations: precision falls to 0.47 ± 0.09, recall to 0.57 ± 0.11 and only 45.3 % of the labels are correct. This observation suggests that either its errors are inherently more difficult to detect, or that asking a model to evaluate its own translations poses specific challenges.

5 Conclusion

In this study, we evaluated the ability of two MT systems commonly used by professional translators to translate LSP texts, specifically abstracts of scientific articles in NLP, using a specific error typology. Our results indicate that these tools still produce translations of insufficient quality, highlighting that MT remains an unsolved challenge in this context. Additionally, we demonstrated that LLMs can be leveraged to identify errors in LSP translations. Future research should explore other domains and languages and assess whether and to what extent automatic error detection can assist translators in their daily work.

Limitations

As previously mentioned, our study relies on commercial translation engines, whose underlying models and training data remain largely unknown. This significantly limits the depth of analysis we can perform on these systems' results. This methodological choice is primarily due to the complexity of the required annotations, which demand extensive and time-consuming manual work. Given our limited capacity for annotation, we focused on evaluating tools that professional translators use in their daily work. However, in our second experiment on error identification, we also included freely available models to broaden our analysis.

Moreover, our study is limited to a single language and a single specialised domain, restricting the generalisability of our findings. This choice stems from the difficulty of obtaining high-quality annotations necessary for evaluating translations in specialised contexts.

Finally, our experiments were conducted on a computer equipped with an NVIDIA A100 GPU with 40GB of memory, except for those involving ChatGPT-4o, which were performed via OpenAI's API. We estimate that executing all our prompts required a total of less than 5 computing hours.

References

Guy Aston. 1999. Corpus use and learning to translate. *Textus, 12(1)*, pages 289–314.

Houda Bouamor and Hassan Sajjad. 2018. H2@bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

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

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2011. Designing a Learner Translator Corpus for Training Purposes. In Natalie Kübler, editor, *Corpora, Language, Teaching, and Resources : From Theory to Practice. Bern: Peter Lang*, volume Etudes Contrastives of *Corpora, Language, Teaching, and Resources : From Theory to Practice. Bern: Peter Lang*, pages 221–248. Peter Lang.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, London.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages

1066–1083, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrst-

edt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sui He. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Natalie Kübler. 2008. MeLLANGE Final Report. Intern report, Université Paris Diderot.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Margaret Rogers. 2015. *Specialised Translation: Shedding the 'Non-Literary' Tag*. Palgrave Macmillan London.

Sai Cheong Siu. 2023. Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation. *SSRN Electronic Journal*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

A Error typology

We have reproduced in Figure 2 the typology of errors used to annotate our data. Errors specific to LSPs are shown in red. In the final version, we will include a link to the full annotation guide, which provides a detailed description and examples of each error (the link cannot be provided at this time to respect the anonymity of the submissions).

B Prompts

Figure 3 describes the full prompt (in French!) for ChatGPT to locate and categorise errors in our corpus. The simpler prompt we used to ask all LLMs to locate errors (without categorising them) is given in Figure 4

C Score Distribution

We have represented in Figure 5 the distributions of precision, recall and F_1 scores across documents.

- Error Typology
  - Content transfer
    - Omission ..... TR-OM
    - Addition ..... TR-AD
    - Distortion ..... TR-DI
    - Indecision ..... TR-IN
    - Source-language-intrusion
      - Untranslated-translatable ..... TR-SI-UT
      - Too-literal ..... TR-SI-TL
      - Units-of-weight-measurement-dates-numbers ..... TR-SI-UN
    - Target-language-intrusion
      - Translated DNT ..... TR-TI-TD
      - Too-free ..... TI-TF
  - Language
    - Syntax ..... LA-SY
      - Determiners ..... LA-SY-DET
      - Wrong-preposition ..... LA-SY-PR
      - Complex-NP ..... LA-SY-GNC
    - Inflection-agreement
      - Tense-aspect-voice ..... LA-IA-TA
      - Gender ..... LA-IA-GE
      - Number ..... LA-IA-NU
    - Typography
      - Spelling ..... LA-HY-SP
      - Accents-diactritics ..... LA-HY-AC
      - Incorrect-case-upper-lower ..... LA-HY-CA
      - Punctuation ..... LA-HY-PU
    - Register
      - Inconsistent-with-ST ..... LA-RE-IS
      - Inadequacy-for-TT ..... LA-RE-IT
    - Style
      - Awkward ..... LA-ST-AW
      - Tautology ..... LA-ST-TA
      - Title-style ..... LA-ST-TS
    - Unclear-reference ..... LA-UR
    - Textual-conventions
      - Coherence ..... LA-TC-CE
      - Cohesion ..... LA-TC-CN
    - Terminology-and-lexis
      - Incorrect-choice-terminology ..... LA-TL-INS
      - Incorrect-choice-lexis ..... LA-TL-ING
      - Incorrect-abbreviation-acronym ..... LA-TL-MAA
      - False-cognate ..... LA-TL-FC
      - Term-translated-by-non-term ..... LA-TL-NT
      - Inappropriate-collocation-SP ..... LA-TL-ICS
      - Inappropriate-collocation-GL ..... LA-TL-ICG
      - Inconsistent-with-TT ..... LA-TL-IT
      - Terminological-inconsistency
        - Different-terms-in-translation ..... LA-TL-TI-DT
        - Different-abbreviations-in-translation ..... LA-TL-TI-DA
  - Tools
    - Hallucination ..... OU-TAH
    - Corpus-conformance ..... OU-CC
    - Duplication ..... OU-DU
    - Incompatible-with-glossary ..... OU-GC

Figure 2: The error typology used in our experiments.

1. Tâche : annoter une traduction
   Objectif : repérer des erreurs sur la base d'une typologie d'erreurs
   ↪ que je te fournis.
   Type de texte : résumé d'article scientifique dans le domaine du TAL
   Fichier joint : MANUEL D'ANNOTATION, qui contient des explications plus
   ↪ détaillées et des exemples des types d'erreurs que je vais te
   ↪ fournir ci-dessous.
   Présentation de la sortie :
   - 1re phrase source
   - 1re phrase cible dans la traduction
   - liste les erreurs
   Etc. jusqu'à la fin de la traduction

   Je vais te donner la typologie d'erreurs.

2. Typologie d'erreurs à suivre méticuleusement : veille à utiliser les
   ↪ types d'erreurs présents et n'en invente aucun. De même, respecte
   ↪ les codes liés à chaque type d'erreur à la lettre ; ne prends donc
   ↪ aucune liberté.
   Explication de la typologie : elle est divisée en 3 grandes catégories
   ↪ d'erreurs : les erreurs de transfert de contenu (erreurs altérant
   ↪ le sens du message ou entravant sa compréhension), les erreurs de
   ↪ langue, et les erreurs liées aux outils ou à leur maîtrise.
   Voici la typologie :
   1. Transfert-contenu (GRANDE CATÉGORIE, NE PAS UTILISER)
   1.1. Omission_TR-OM
   * Une omission se produit lorsqu'il manque, dans la traduction, une
   ↪ idée qui est présente dans le texte source. Il ne faut pas
   ↪ confondre omission et implicitation. Une omission a lieu sans
   ↪ réelle raison valable, alors qu'une implicitation est un moyen
   ↪ d'éviter une surtraduction.
   1.2. Rajout_TR-AD
   * À l'instar de la différence entre omission et implicitation, on peut
   ↪ souligner une différence de nuance entre le rajout et
   ↪ l'explicitation. L'ajout est considéré comme une erreur, alors que
   ↪ l'explicitation peut s'expliquer par le fait que le traducteur ou
   ↪ le post-éditeur souhaite éviter la sous-traduction.
   ... jusqu'au bout de la typologie ...

   - Prête attention à tous les aspects, autant le transfert de contenu
   ↪ que la langue et la terminologie et les erreurs liées aux outils.
   - Si tu as besoin d'exemples, réfère toi au manuel d'annotation en
   ↪ pièce jointe.

   Je vais te donner la traduction à évaluer avec son texte source.

3. Voici le texte source et sa traduction à annoter :
   (source text)
   (target text)

   PROCÈDE À L'ANNOTATION. Attention, n'annote QUE les erreurs, pas des
   ↪ améliorations ou suggestions ! Il peut y avoir plusieurs erreurs
   ↪ dans une même phrase.

Figure 3: Prompt used on GPT-4o

```
Your task is to annotate a translation by identifying the
errors it contains.

I will give you abstracts of scientific articles written in
English that are translated into French.

Pay attention to all aspects, from content transfer to
language and terminology and tool-related errors.
----------
Here is the source text:

{{source_text}}

and here is its translation that you must annotated:

{{target_text}}
----------
There may be several errors in the same sentence. Your
answer must be a JSON list containing the list of errors.
Each error must be described by a dictionary with the
following key:
- "span": a substring of the translation that indicates
the words in the translation that must be corrected for
the translation to be correct and nothing else.
You must identify the smallest span possible.
- "beginning": the position of the error in the translation,
described as the index of the first character in the
translation.

For each error only select the smallest span possible.
Return only the JSON and do not include any other
information or comment.

PROCEEDS WITH THE ANNOTATION. Please note that you should
ONLY annotate errors, not propose improvements or
suggestions!
```

Figure 4: The “small” prompt we used for error identification.

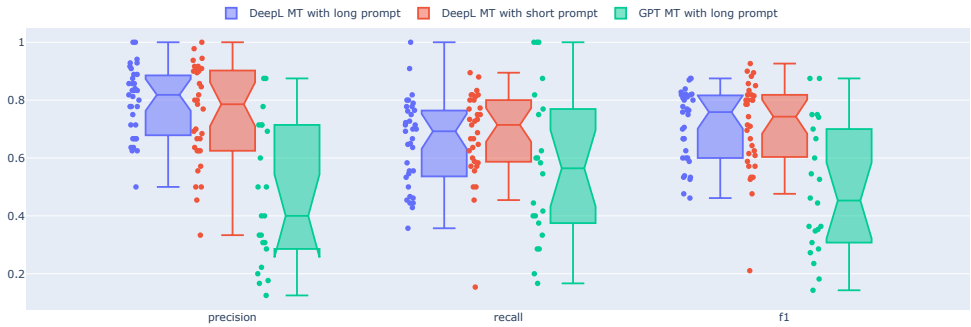

Figure 5: Distribution of metrics across documents for the different prompts we consider.