Fine-tuning NMT Models and LLMs for Specialised EN-ES Translation Using Aligned Corpora, Glossaries, and Synthetic Data: MULTITAN at WMT25 Terminology Shared Task

Lichao Zhu Maria Zimina-Poirot Cristian Valdez Stéphane Patin

ALTAE, Université Paris Cité
{lichao.zhu, maria.zimina-poirot, cristian.valdez, stephane.patin}@u-paris.fr

Abstract

This paper presents a hybrid evaluation of terminology-aware English-to-Spanish machine translation systems developed for the WMT25 Terminology Shared Task, specifically targeting the Information Technology (IT) domain. Our objective was to improve terminology accuracy and overall translation quality and highlight the potential of specialised terminology-aware translation models for technical domains. We used different enhancement strategies for both neural machine translation (NMT) systems and large language models (LLMs). These strategies include fine-tuning with synthetic data, the use of in-domain parallel corpora, and hard constraint methods such as placeholder substitution and in-context glossary integration. The results demonstrate distinct lexical and stylistic profiles in the outputs of fine-tuned NMT systems and LLMs, as well as the complementary advantages of different terminology injection methods. Systems behave differently with and without a glossary, as demonstrated by experimental results. The NMT systems seem to be rather limited in adapting to special lexicons and resizing embeddings, which is the opposite of LLMs, which prefer structured instructions. Although our translation systems achieved their highest scores on the *NoTerm, Consistency* metrics, exceeding 81%, demonstrating their ability to produce stable and coherent translations of recurring terms and phrases in unconstrained settings, the precision of the terminology and overall quality of the translation could have been improved by additional training.

1 Introduction

Adaptation of the NMT model to a specific domain (Chu et al., 2017; Chu and Wang, 2018; Saunders, 2021) is a major concern in bilingual neural machine translation. Among the various methods that have been proposed to tackle domain adaptation, two approaches are particularly relevant to the objective of this shared task: i) **Data approach** that involves selecting and filtering existing in-domain parallel segments (Moore and Lewis, 2010; Axelrod et al., 2011), or generating synthetic data. The latter is widely used in back-translation and in enhancing domain-specific data by reformulating or paraphrasing (Sennrich et al., 2016; Edunov et al., 2018); ii) **System approach** which aims to assign weights to segments close to the target domain (Wang et al., 2017). In recent years, research has increasingly focused on frugal domain adaptation strategies, emphasising the efficient use of limited resources and optimising training settings to address low-resource scenarios (Adams et al., 2022; Marashian et al., 2025). With the rapid development of LLMs for translation purposes in recent years, several methods for LLM domain adaptation have emerged, such as prompt engineering and context leaning (Zhang et al., 2023; Pourkamali and Sharifi, 2024; Yamada, 2023), constrained decoding to enforce terminology or special data format (Luca Beurer-Kellner, 2024; Bogoychev and Chen, 2023), Supervised Fine-Tuning (SFT) using reinforcement learning from human feedback (Ouyang et al., 2022), etc. Meanwhile, domain adaptation sheds light on the functioning, strengths, and weaknesses of LLM (Lu et al., 2025a), making these 'black boxes' more interpretable – an important objective of our participation in the shared task. In our contribution, we have chosen to use three strategies to fine-tune both NMT and LLM systems to ensure accurate translation of terms tagged as 'proper_terms' in the dev set:

- An open source NMT system was fine-tuned by employing placeholders for lexical-constrained decoding using additional aligned segments within the domain.
- An artificially augmented training data set

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1042–1049
November 8–9, 2025.

was created using a prompt system and used to fine-tune a baseline on a commercial model training server.

- Using aligned segments and a glossary, Low-Rank Adaptation was used to make minor changes to an LLM (Hu et al., 2021).

2 Data processing and augmentation

2.1 Data sources for domain-specific model specialisation

To achieve precise terminology and consistency in fields like finance, IT and legal texts, it is essential to use high-quality aligned corpora and domain-specific glossaries to fine-tune NMT models and LLMs for specialised machine translation. The use of synthetic data generation methods can also help to augment domain-specific corpora, enhancing models' abilities to manage specialised terminology and contexts effectively. Prompt-based generation, retrieval-augmented generation, self-instruction, and reinforcement learning with feedback are some of the approaches available for synthetic data generation (Lu et al., 2025b; Nad et al., 2025). These methods can improve model performance, data diversity, and adaptability to domain-specific requirements by supplementing real training data with synthetic examples generated by LLMs.

2.2 High-quality parallel corpus

To increase the size of the training data and achieve a broader terminological coverage, we used a high-quality parallel corpus in IT and closely related fields: the European Union Intellectual Property Office (EUIPO) Trademark and Design Guidelines in the production, technology and research domains, translated by professional translators from English into Spanish.[1] This resource, created by the European Language Resource Coordination (ELRC), contains 16,439 translation units; 386,472 tokens in English and 424,702 tokens in Spanish. After filtering based on length control and the basic alignment quality of the segments, we obtained 6,359 parallel segments containing 62,481 English tokens and 67,791 Spanish tokens, representing an average of 910 words (60-66 characters) per segment.

[1] https://is.gd/LlPIYr

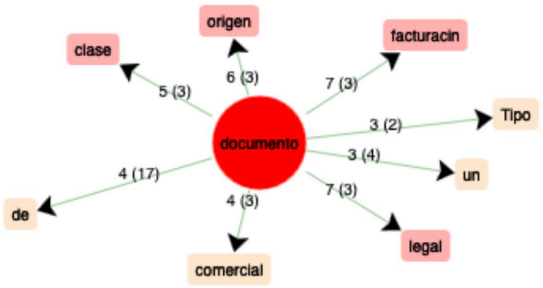

Figure 1: Parallel co-occurrences in the augmented data, measuring the frequency and specificity of lexical attractions.[2]

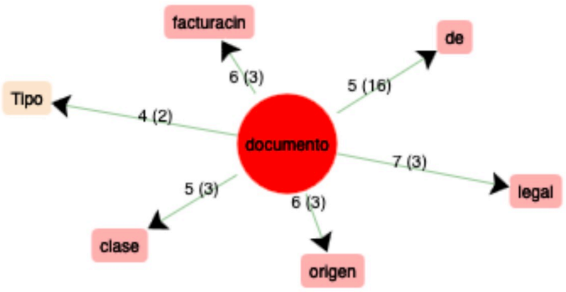

Figure 2: Parallel co-occurrences in the development set of the shared task, measuring the frequency and specificity of lexical attractions.

2.3 In-domain synthetic data

In our experiments, we used the dev data set (enes-dev: 500 aligned segments) augmented via a system of prompts on ChatGPT-4 (first to generate an initial data set in English using task-specific terminology) and on DeepSeek (to translate the generated dataset into Spanish using the WMT 2025 dev set term correspondences). We used free public versions of both platforms. Our objective was not only to consider specific term correspondences, but also to capture linguistic features of the test set to create synthetic data reproducing the dev set patterns. We created a synthetic dataset enes-AUG (1,746 aligned English-Spanish sentence pairs, 10,021 tokens in English, 10,822 tokens in Spanish). The generated synthetic data set has some recognisable characteristics in terms of regularly repeated characteristic patterns (for example, in Spanish: *Usa las acciones **en tu documento**. Utilice los Servicios profesionales **en su documento**. Utilice el Soporte SAP **en su documento***).

[2] Calculated by iTrameur (textometrics tool): https://itrameur.clillac-arp.univ-paris-diderot.fr

We analysed parallel co-occurrence networks (Zimina and Fleury, 2016) in sentences containing the same terms in the dev set and in the augmented data. It reveals that both data sets share key terminological traits (see Figures 2.3 and 2), supporting the suitability of AI-generated data for our experiment. However, the AI-generated data set lacks the uppercase usage typical of the enes-dev data. For example, the term "Source Document Category" appears in lowercase. Future experiments could include prompt instructions to better replicate capitalisation in terminology.

2.4 Glossary terms and augmented terminology

The enes-dev set contains some rather challenging (or even questionable) annotations of "proper_terms" and "random_terms", for example:

"en": **"Source Document Category"**, "es": **"Tipo de documento de origen"**, "proper_terms": {"source document category": "tipo de documento de origen"}, "random terms": {"Source": "Origen", "Document": "documento", "Category": "tipo"}.

If ***Tipo de documento de origen*** is a term, then its constituents, such as ***documento*** and ***tipo*** are hardly random in specialised discourse, especially if we consider term variation, which involves using different forms to express the same or nearly the same concepts within a specialised domain (Daille, 2017). In our work, we considered that such occurrences are part of specialised discourse and their distributional properties (Wingfield and Connell, 2022) are reflected in the augmented dataset. We also tried to take into account the fact that different multiword terms possibly contain common lexical items. In this respect, term variation is reflected in the augmented data set. For example, while the term 'source document category' is translated by *tipo de documento de origen*, our synthetic data set also contains contexts, where 'document type' is translated by *clase de documento*:

en: Save the **source document category**. > es: *Guarda el **tipo de documento de origen***.

en: Save the **document type**. > es: *Guarda la **clase de documento***.

3 Systems

3.1 NMT system

We investigated methods to impose lexical constraints on NMT systems that do not inherently support glossary use. One approach involves fine-tuning a generic NMT model with domain-specific synthetic data generated through LLMs and other generative AI tools. This synthetic data set incorporates targeted terminology to guide system translations, improving sentence level accuracy and ensuring consistent terminology usage. The fine-tuning process leverages data augmentation techniques to integrate specific term correspondences and improve translation coherence within specialised domains.

3.2 LLM

Recent research suggests that LLMs can be effectively tuned for MT using surprisingly small amounts of high-quality parallel data. Fine-tuning with LLMs (such as Llama-2 7B) can deliver strong performance when trained on as few as 32,219 parallel sentences (Lu et al., 2025a). This is in line with the hypothesis of "superficial alignment", which suggests that LLMs have already mastered their translation skills during pre-training, and fine-tuning concentrates on aligning the model with the specific task format.

4 Experimental settings

We used constrained machine translation with hard terminology control (EN-ES) to ensure that English terms were consistently predicted as their Spanish equivalents. We implemented our NMT system with Marian (Junczys-Dowmunt et al., 2018), mainly because it allowed constraint decoding with the additional_special_tokens parameter, and the models embeddings were resized accordingly. In the training data, we replaced English terms and their Spanish equivalents with placeholders: each pair of bilingual terms was replaced by the same placeholder. Consequently, English terms in the input were replaced by placeholders before decoding, and placeholders in the output were replaced by the equivalent Spanish terms after generation. Our system was fine-tuned with Seq2SeqTrainer with the following setting: training epochs: 5, learning rate: 1e-5, training batch size: 8, gradient accumulation steps: 2, max length: 128.

For the LLM model, we used EuroLLM-1.7B-Instruct (Martins et al., 2024) as a causal LM due to its relatively light weight for fine-tuning. For each pair of aligned segments (EN-ES), the system builds a prompt that injects the glossary before instruction as follows:

```
"### Instruction:"
"Translate the following text from English
to Spanish using the provided glossary."
"### Glossary (Information Technology):"
"{glossary_items}"
"### Text in English:"
"{src}"
"### Translation in Spanish:"
"{tgt}"
```

The glossary was created using the English and Spanish lexicons tagged as "proper_term" elements in the dev set. After filtering and deduplication, we obtained 172 unique occurrences in English (not case-sensitive). These occurrences correspond to 221 pairs of English and Spanish terms. The training was set as follows: `training batch size`: 4, `learning rate`: 1e-4, `weight decay`: 0.01, `lr scheduler type`: linear, `warmup steps`: 100, `training epochs`: 1. We used parameter-efficient LoRA (Hu et al., 2021) adapters for light fine-tuning, using 8-bit loading to reduce GPU memory usage: `task type`: CAUSAL LM, `r`: 16, `lora alpha`: 32, `lora dropout`: 0.1. For both the Marian and EuroLLM fine-tuning processes, we used a small number of epochs and a small batch size to reduce training time and prevent over-fitting. The main purpose was to evaluate the efficiency of our training pipelines.

We also used an advanced commercial MT platform SYSTRAN Model Studio Lite[3] to fine-tune a generic EN-ES baseline model. NMT was used to ensure accurate sentence-level translations. GenAI, including LLMs, was used to develop a synthetic data set that contains domain-specific terms to fine-tune the initial system and implement specific term correspondences, improving coherence regarding terminology specifications (see Section 2.3 on data augmentation techniques).

[3] https://modelstudio-lite.systran.net

5 First results

5.1 Quantitative metrics

We investigated lexical characteristics and vocabulary usage in six translations produced by different models (see Table 1) and compared them using quantitative methods, such as `correspondence analysis`, `vocabulary growth assessment`, and `characteristic elements computation` (Lebart et al., 1998). The investigation revealed nuanced differences in the translations and highlighted the impact of glossary inclusion and placeholder handling on translation quality and style.

A corpus aggregated from six different translations contains a total of 35,250 tokens. On average, each translated text comprises approximately 5,875 tokens, with two notable exceptions: **EuroLLM_SEG_Glossary** (6,192 tokens) and **MTwithplaceholder** (5,400 tokens).

5.1.1 Vocabulary growth

The vocabulary growth curves, which reflect the natural increase in distinct words encountered, show that **EuroLLM_SEG_Glossary** consistently has the largest vocabulary size (see Figure 3). This suggests that it has the most diverse vocabulary of all the models tested. **Systran_SEG_AUG** exhibits somewhat lower vocabulary growth, very close to the reference translation, while **Marian_SEG** shows the slowest increase, suggesting a smaller or more limited vocabulary compared to the others.

5.1.2 Characteristic elements

The results of `characteristic elements` computation reveal the use of English lexical units in **MTwithplaceholder**, which do not correspond to common Spanish borrowings, highlighting the models lexical weaknesses. **EuroLLM_SEG_Glossary** distinctly exhibits a preference for informal address forms, favouring pronouns and verbs such as *puedes*, *selecciona*, and *tus* over their formal counterparts like *pueda* or *Seleccione*. It also underuses the term *plantilla*, substituting it with *modelo* or omitting it entirely, indicating stylistic or glossary-driven variations[4].

[4] The translations of our systems are available at https://github.com/lichaozhu/MULTITAN-WMT2025-Terminology

[5] Generated with iTrameur (textometrics tool): https://itrameur.clillac-arp.univ-paris-diderot.fr

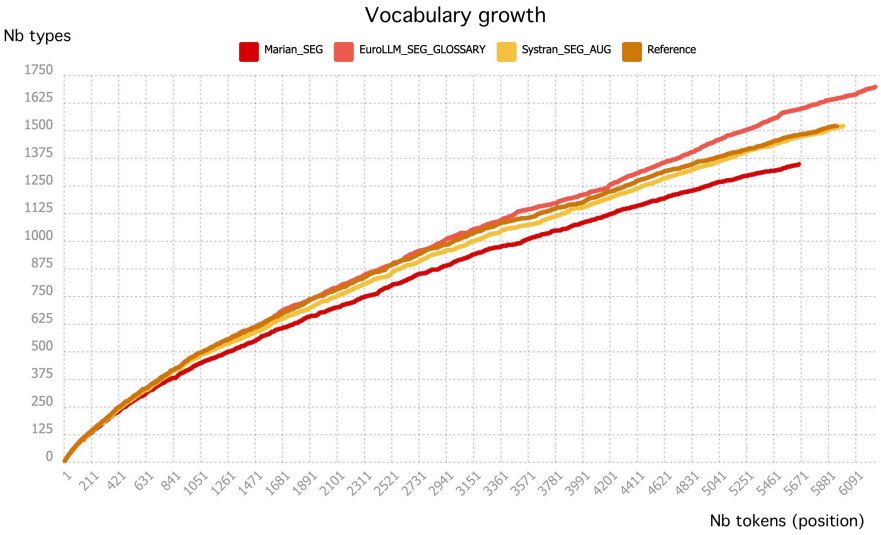

Figure 3: Comparison of vocabulary growth curves among submitted translations and the reference.[5]

5.1.3 Correspondence analysis

`Correspondence analysis` provided a multidimensional visualisation of the lexical relationships across the models (see Figure 4). The first factorial plane, which accounts for 49.9% of the total lexical variation, highlights the following oppositions:

- **(Pure) MT translations vs. (human) reference**: As expected, the results of the correspondence analysis highlight the differences between the reference translation and the translations produced by the submitted MT systems.
- **NMT vs. LLM**: Submitted MT translations are positioned along a continuum that differentiates between NMT and LLM characteristics. The **Systran_SEG_AUG** system, which relies on LLM-generated data, exhibits hybrid traits.

The second factorial plane (see Figure 4), which accounts for 26.7% of the total lexical variation, further confirms oppositions and proximities observed in the first plane, mapping a closer proximity between **Systran_SEG_AUG** and the reference translation compared to the other submitted translations:

- **Systran_SEG_AUG** shows some characteristics that approximate the **reference** translation.

5.1.4 Performance scores

Following this analysis, we conducted an evaluation using the reference translations provided by the organisers. Table 2 compares the performance of pre-trained or generic models (Marian MT, EuroLLM, and Systran *Generic*) with that of fine-tuned models. Our analysis shows that although constrained decoding helps models better translate "proper terms", it reduces the overall translation performance of MT systems. This effect is particularly evident in Marian MT: although the fine-tuned model translated more proper terms and produced more formal equivalences, its overall translation quality decreases, with the COMET-DA score dropping from 0.86 (pre-trained) to 0.82 (fine-tuned). However, constrained decoding improves the term-matching precision of EuroLLM without heavily degrading its overall translation quality. For Systran, the score difference between the pre-trained and fine-tuned models is minimal.

5.2 Linguistic analysis

Evaluating fluency is difficult since most segments are decontextualised noun phrases, but there are examples where fluency differences between models are noticeable (see Table 3).

Systran_SEG_AUG tends to produce translations that are quite literal but accurately preserve the source content. The system generally yields the best results in terms of information transfer. **Marian_SEG** model sometimes omits segments. When this happens, there are negative effects on orthographic correctness, such as missing capital

[6]Generated with **Lexico 5**, lexicometrics software: https://lexi-co.com/L5Presentation.html

[7]*: Submitted version.

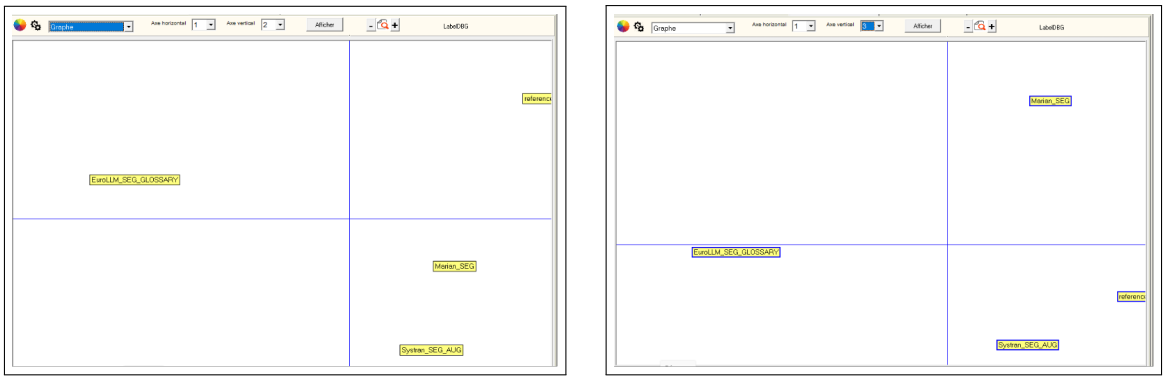

Figure 4: Factorial correspondence analysis of submitted translations. The first factorial plane, formed by axes 1 and 2, accounts for 49.9% of the variation. The second one, formed by axes 1 and 3, accounts for 26.7%.[6]

System	Token	Form	Hapax	Fmax
ES_SystranIT	6,014	1,519	834	653
ES_Systrangeneric	5,976	1,504	849	677
EuroLLM_SEG_GLOSSARY*[7]	**6,192**	**1,670**	**990**	**690**
MTwithplaceholder	5,400	1,592	973	440
Marian_SEG*	5,664	1,351	762	658
Systran_SEG_AUG*	6,004	1,524	866	684
Total	35,250	2,871	873	3,802

Table 1: Quantitative characteristics of the translations generated by each of the models.

letters at the beginning of sentences.

EuroLLM_SEG_GLOSSARY model occasionally leaves some English segments untranslated. This occurs rarely but more often than with other models. Some untranslated segments result from poor segmentation in the source text. There is no consistent correlation between the omission or retention of expected terms across models. For example, sometimes expected terms are omitted, sometimes other segments are omitted leaving only the expected term, and in some rare cases the source text remains unchanged.

In cases where **EuroLLM_SEG_GLOSSARY** leaves English segments untranslated (e.g., "On the To Be Billed"), it appears to stem from improper segmentation. For example, the segment "On the To Be Billed" that is left in the translation generated by **EuroLLM_SEG_GLOSSARY** is the result of poor segmentation of the original version. If the decision was to leave the tab name in English, it should have been rendered as "*en la pestaña To Be Billed*". **EuroLLM_SEG_GLOSSARY** sometimes introduces lexical creations or barbarisms, which can make understanding difficult.

In summary, **Systran_SEG_AUG** performs best at preserving source information. **Marian_SEG** suffers from segment omission causing orthographic errors. **EuroLLM_SEG_GLOSSARY** tends to leave English untranslated more often and sometimes produces confusing lexical forms.

5.3 Interpretative insights

This combined quantitative and qualitative analysis shows that translation models vary in more than just lexical richness; they also differ in terms of stylistic choices and lexical specificity. There are significant differences in how models handle formality, lexical borrowing, and terminology consistency, reflecting variations in MT architecture and preprocessing (e.g. glossaries and placeholders). Correspondence analysis is a useful tool for visualising these differences, as it reduces complex lexical data into interpretable axes of variation. This enables more informed evaluations of MT output.

6 Conclusion and perspectives

Our contribution to the shared task reveals the strengths and challenges of NMT systems and LLMs in translation tasks. Although NMT systems perform well with small amounts of high-quality domain-specific training data, their performance can deteriorate under constrained decoding conditions. In contrast, LLMs can benefit from structured guidance, such as glossaries and clear instructions, which enhances their translation quality. These findings emphasise the complementary

System	Total terms	Translated terms	Ratio	BLEU	chrF	COMET DA
EuroLLM_SEG_GLOSSARY	538	204	37.9%	38.6	69.1	0.83
EuroLLM_1.6B_Pre-trained	538	199	37.0%	36.1	53.4	0.84
Marian_SEG	538	**254**	**47.2%**	48.1	73.2	0.82
Marian_Pre-trained	538	196	36.4%	45.0	58.1	0.86
Systran_SEG_AUG	538	222	41.3%	50.7	**75.7**	**0.88**
Systran_Generic	538	221	41.1%	**51.2**	73.2	0.88

Table 2: Term coverage and evaluation scores

ID	Source text	Systran_SEG_AUG	Marian_SEG	EuroLLM_SEG_GLOSSARY	Term
1	On the To Be Billed Tab, select one or more items as required and choose Write Off.	En la ficha Para facturar, seleccione uno o más elementos según sea necesario y elija Cancelar.	En la pestaña Para ser facturado seleccione uno o más elementos como sea necesario y seleccione Escribir apagado.	En la pestaña On the To Be Billed, seleccione uno o más elementos según sea necesario y elija Deshacer.	Write off<>ignorar
2	On the To Be Billed Tab, select one or more items as required and choose Restrict Date.	En la ficha Para facturar, seleccione uno o más elementos según sea necesario y Restringir fecha.	En la pestaña Para ser facturado seleccione uno o más elementos como sea necesario y seleccione Limitar fecha.	En la pestaña On the To Be Billed, seleccione uno o más elementos según sea necesario y Restrict Date.	Tab<>pestaña

Table 3: Translation comparisons

nature of the two approaches, suggesting that combining NMT's data-driven learning with LLMs' flexible, instruction-driven capabilities could result in more robust and effective translation solutions.

We also identified some limitations through our experimentation. None of our open-source systems translated all the segments from English to Spanish, with some segments always remaining untranslated. This is probably due to gaps in the lexicons of the training and test data, or to forced decoding effects that prioritise and select the source token over the target token. In summary, a trade-off must be made between model complexity, performance and training cost. While a 16B-parameter model could potentially avoid missing any translations, it is much more challenging to fine-tune such a model due to its substantial number of parameters, size, and the associated high training costs in terms of GPU memory, time, and environmental impact.

7 Acknowledgements

This research was funded by the 2024 **MULTITAN-GML** Research Equipment Grant (COPES-2024-12, *Fonds d'intervention Recherche, Université Paris Cité*) and used (**PNS-UP:** plateforme de traduction automatique sur serveur université) scientific platform[8].

References

Virginia Adams, Sandeep Subramanian, Mike Chrzanowski, Oleksii Hrinchuk, and Oleksii Kuchaiev. 2022. Finding the right recipe for low resource domain adaptation in neural machine translation. *Preprint*, arXiv:2206.01137.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In

[8] https://plateformes.u-paris.fr

Proceedings of the 27th International Conference on Computational Linguistics, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Béatrice Daille. 2017. *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ludovic Lebart, Salem André, and Berry Lisette. 1998. *Exploring Textual Data*. Academic Kluwer Publisher, Dordrecht, Boston, London.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2025a. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(84).

Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2025b. Machine learning for synthetic data generation: A review. *Preprint*, arXiv:2302.04062.

Martin Vechev Luca Beurer-Kellner, Marc Fischer. 2024. Guiding llms the right way: Fast, non-invasive constrained decoding. *arXiv*.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2409.16235*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 220–224.

Mihai Nad, Laura Dioan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615134633.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *Preprint*, arXiv:2401.08429.

Danielle Saunders. 2021. *Domain Adaptation for Neural Machine Translation*. Ph.d. thesis, University of Cambridge, Cambridge, United Kingdom. Available online.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Cai Wingfield and Louise Connell. 2022. Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, 37(10):1220–1270.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

Maria Zimina and Serge Fleury. 2016. Perspectives de larchitecture trame/cadre pour les alignements multilingues. *Nouvelles perspectives en sciences sociales*, 11(1):325–353.