LUXINSTRUCT: A Cross-Lingual Instruction Tuning Dataset For Luxembourgish

Anonymous ACL submission

Abstract

Instruction tuning has become a key technique for enhancing the performance of large language models, enabling them to better follow human prompts. However, low-resource languages such as Luxembourgish face severe limitations due to the lack of high-quality instruction datasets. Traditional reliance on machine translation often introduces semantic misalignment and cultural inaccuracies. In this work, we address these challenges by creating a cross-lingual instruction tuning dataset for Luxembourgish without the use of machine translation. Instead, by leveraging aligned data from English, French, and German, we build a high-quality dataset that preserves linguistic and cultural nuances. We provide evidence that this cross-lingual approach not only circumvents common translation pitfalls but also leads to higher cross-lingual alignment within LLMs. This alignment is essential for enabling effective transfer to low-resource languages such as Luxembourgish. Therefore, our results advocate for data curation strategies that prioritize linguistic integrity over automated translation.

1 Introduction

In recent years, instruction tuning has emerged as a crucial technique in the development of Large Language Models (LLMs), significantly enhancing their ability to follow user prompts across a wide range of tasks. By fine-tuning models on curated datasets of instruction-output pairs, LLMs have been enabled to generalize better, respond more naturally, and align more closely with human intent (Ouyang et al., 2022). However, despite its success in high-resource languages, instruction tuning remains a significant challenge for low-resource languages. One of the key bottlenecks is the scarcity of high-quality instruction-following datasets in these languages. Unlike English, where vast corpora of annotated instructions are available, many low-resource languages lack sufficient data, both in quantity and in variety, to effectively fine-tune LLMs. The process of manually creating instruction datasets is labor-intensive and expensive, often requiring native speakers with expertise in both the language and various task domains. Consequently, researchers have frequently resorted to machine translation (MT) techniques to generate instruction data for these languages (Li et al., 2023; Holmström and Doostmohammadi, 2023; Li et al., 2024). However, relying on MT to produce instruction tuning data introduces several complications. Translations may fail to capture the nuanced meanings, cultural contexts, and idiomatic expressions inherent in the source language, leading to instruction-response pairs that are misaligned or unnatural in the target language (Bizzoni et al., 2020). This misalignment can adversely affect the performance of LLMs trained on such data (Yu et al., 2022), as they may learn to generate responses that are semantically incorrect or culturally inappropriate.

A language that exemplifies these challenges is Luxembourgish, a West Germanic language with about 400 000 speakers in Luxembourg. As a low-resource language, it suffers from a paucity of linguistic data, making it difficult to develop robust language or MT models.

To address the scarcity of high-quality data, we compile LUXINSTRUCT, a cross-lingual instruction tuning dataset for Luxembourgish. By avoiding machine translation into Luxembourgish, our approach preserves linguistic integrity, while enabling the adaptation of LLMs to Luxembourgish through alignment with English, French, and German. Additionally, the use of human-generated, rather than synthetic, data guarantees LUXINSTRUCT's high quality.

Our findings indicate that this cross-lingual dataset, due to its native construction, offers superior quality and has the potential to result in more effective fine-tuning outcomes compared to monolingual MT-based instruction tuning data.

2 Related Work

2.1 Luxembourgish NLP

Luxembourgish NLP is still in its early developmental phase. The field gained traction with the introduction of the encoder-only model LUXEMBERT (Lothritz et al., 2022), followed by the decoder-only LUXGPT-2 (Bernardy, 2022), and later the encoder-decoder models LUXT5 and LUXT5-GRANDE (Plum et al., 2025). Nonetheless, Lothritz and Cabot (2025) demonstrated that both open-source and many proprietary LLMs still fall short of achieving high-level performance in Luxembourgish.

In terms of existing datasets, the most substantial compilation of unlabeled Luxembourgish text to date was assembled by Plum et al. (2025), while Philippy et al. (2025) contributed a parallel corpus covering English–Luxembourgish and French–Luxembourgish pairs. Nevertheless, a native high-quality instruction tuning dataset has yet to be developed.

2.2 Low-Resource Language Instruction Tuning Data

While prior work has focused on creating instruction tuning datasets for specific languages (Suzuki et al., 2023; Azime et al., 2024; Laiyk et al., 2025; Shang et al., 2025), many languages—including Luxembourgish—still lack such resources. This gap is largely due to the high cost of manually curating instruction tuning data for low-resource languages. Existing approaches typically rely on machine translation (Li et al., 2023; Holmström and Doostmohammadi, 2023; Li et al., 2024) or repurposing labeled NLP datasets (Muennighoff et al., 2023). However, neither method is effective for Luxembourgish, due to limited translation quality and a scarcity of labeled data.

Köksal et al. (2024) propose the use of *reverse instructions* to generate instruction tuning data from raw text, a method later expanded to to the *Multilingual Reverse Instructions* (MURI) framework (Köksal et al., 2024). Yet, MURI still relies on two rounds of translation and focuses on multilingual (same-language) rather than cross-lingual (instruction and output in different languages) tuning. While multilingual tuning benefits low-resource settings (Weber et al., 2024; Shaham et al., 2024), cross-lingual tuning has been shown to offer comparable advantages (Li et al., 2024; Chai et al., 2024; Lin et al., 2025).

3 LUXINSTRUCT

3.1 Dataset Creation

We create cross-lingual instruction tuning data for Luxembourgish (Figure 1) using three primary data sources: Wikipedia, news articles, and an online dictionary. More information on the source data and the process in provided in Appendix A.

Wikipedia We adopt a reverse instruction generation approach inspired by the MURI framework (Köksal et al., 2024), but diverge in key aspects to avoid translation artifacts. Instead of translating existing instruction data, we prompt OpenAI's `gpt-4.1-mini` to select informative extracts from Luxembourgish Wikipedia articles and generate corresponding instructions directly in English. This allows for high-quality, semantically aligned instruction–output pairs without relying on machine translation. Additionally, unlike MURI, which applies a single prompt to full, often noisy documents, our method ensures cleaner inputs by allowing the model to select coherent spans. Generated pairs are further filtered based on a series of heuristic-based filtering steps (length, correct language, extraction consistency, etc.) to ensure data quality. In order to expand the multilinguality and utility for future research, we additionally machine-translate a subset of the instructions to German, French and Luxembourgish[1]. The resulting dataset forms the **`Open-Ended`** portion of LUXINSTRUCT.

News Articles The Luxembourgish news platform *RTL Luxembourg*[2] publishes articles in Luxembourgish as well as French and English. Since there is no direct alignment between language versions, we use OpenAI's `text-embedding-3-small` model to retrieve bilingual article pairs (LB-EN & LB-FR). From these parallel news articles we then create instruction–output pairs in two different task styles: (1) generating Luxembourgish news headlines from English or French articles (**`Article-To-Title`**), and (2) generating hypothetical Luxembourgish news articles from English or French headlines (**`Title-To-Article`**). Additionally, similar monolingual Luxembourgish instruction-output pairs are created. To support diversity in the instruction phrasing, we employ a set of predefined templates, randomly selected per instance.

[1] We use `gpt-4.1-mini` for French and German, and `gpt-4.5` for Luxembourgish.

[2] https://www.rtl.lu

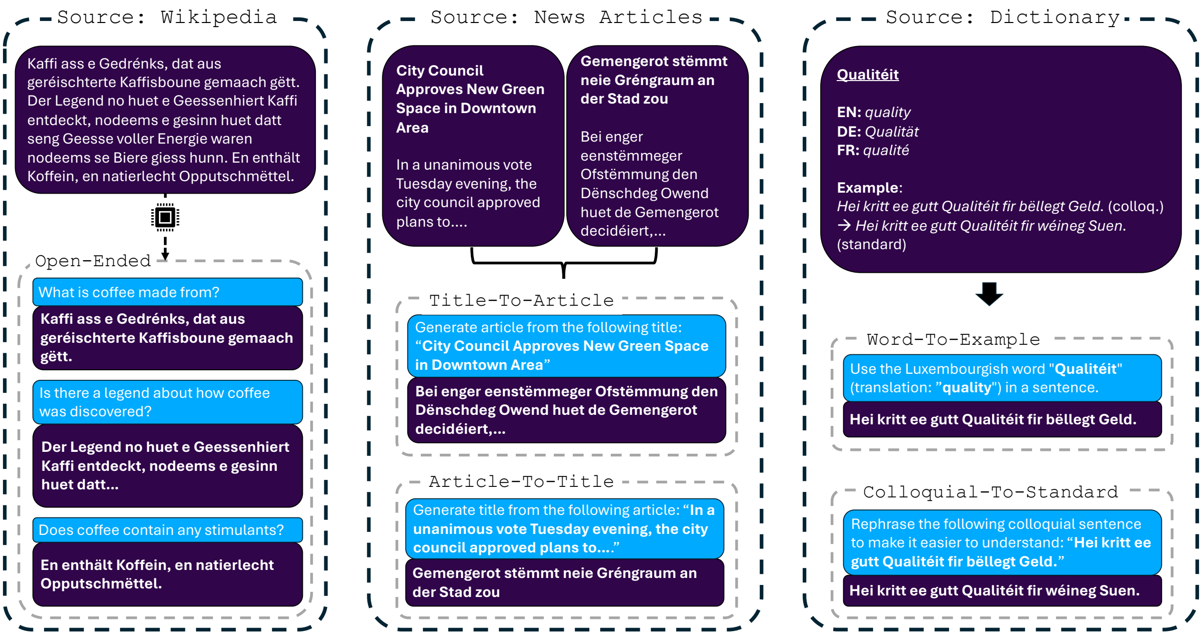

Figure 1: **Overview of the LuxInstruct Data Creation Pipeline.** The dataset is constructed from three core sources: (1) Using **Wikipedia**, we apply a reverse instruction generation approach where a language model extracts informative spans from Luxembourgish articles and directly generates corresponding instructions in English (`Open-Ended`). (2) From **News Articles**, we leverage parallel multilingual content to create cross-lingual instruction–output pairs for both headline generation (`Article-To-Title`) and article generation (`Title-To-Article`). (3) Using an **Online Dictionary**, we design tasks based on lexical entries and example sentences, focusing on word usage (`Word-To-Example`) and colloquial sentence simplification (`Colloquial-To-Standard`).

Online Dictionary We leverage a publicly available Luxembourgish dictionary containing lexical entries with translations (to English, French, German) and example sentences. We design two task types: (1) generating Luxembourgish example sentences of a given Luxembourgish word, where the exact word meaning is given by the translation of the word (**`Word-To-Example`**), and (2) simplifying colloquial Luxembourgish sentences (**`Colloquial-To-Standard`**). Again, for both tasks, multiple instruction templates are used to introduce variation in phrasing.

3.2 Dataset Statistics

Our new dataset consists of 277,261 cross-lingual instruction-output samples across English, French, and German as instruction languages, along with 161,564 monolingual samples in Luxembourgish, where both instruction and output are in Luxembourgish. Although the underlying seed data is similar across instruction languages—leading to a high degree of overlap in outputs—we still count 223,913 unique Luxembourgish output strings across all language subsets. Appendix A provides the exact number of samples per language and type of task (Table 1) as well as examples (Table 2)[3]

4 Cross-Lingual Vs Monolingual Instruction Tuning

We conduct a small-scale study to evaluate the impact of cross-lingual instruction tuning data compared to monolingual data.

Due to the lack of robust evaluation resources for Luxembourgish text generation[4], and the unreliability of reference-free methods like *LLM-as-a-judge*—given the current limitations of state-of-the-art LLMs in Luxembourgish—we restrict our evaluation to internal measures of cross-lingual alignment within the model. Such alignment is a crucial prerequisite for effective cross-lingual transfer in LLMs (Gaschi et al., 2023; Wang et al., 2024), particularly benefiting low-resource languages such as Luxembourgish.

Experimental Setup We fine-tune models on a subset of the `Open-Ended` portion of our dataset, chosen for its partially parallel content across four

[3]A larger subset of our dataset is provided here.

[4]While Plum et al. (2025) introduce a valuable four-task benchmark, overlap in seed data between their benchmark and our dataset raises potential data leakage concerns, so we refrain from using it.

languages. Each model is fine-tuned separately using instructions in a single language—English, French, German, or Luxembourgish—while responses are consistently in Luxembourgish.

We then assess the alignment between the Luxembourgish embedding space and the English, French, and German spaces by using parallel data from FLORES-200 (Team et al., 2022) and computing Centered Kernel Alignment (CKA) scores (Kornblith et al., 2019) using the model's mean-pooled hidden states of its last layer. More technical details and information on the used models is provided in Appendix B.

Figure 2 shows the average increase in alignment between the Luxembourgish and the English, French, and German representation spaces after fine-tuning with different instruction languages. While alignment gains vary across models, cross-lingual instruction tuning proves at least as effective—and often more so—than monolingual tuning. EN-LB and FR-LB configurations yield the highest alignment improvements, whereas DE-LB performs often worse than monolingual (LB-LB) tuning. This suggests that pairing low-resource languages with more distant languages during instruction tuning may be more effective than using closely related ones. We provide the exact results per language pair in Table 3 in the appendix.

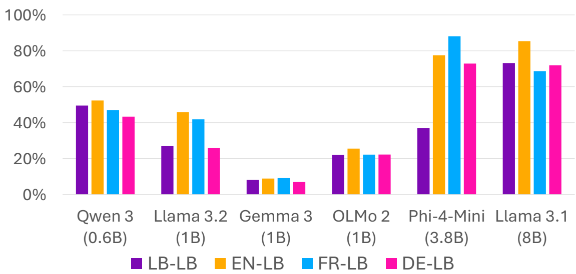

Figure 2: Mean variation (in %) in alignment between the Luxembourgish and the English, French, and German representation spaces after fine-tuning on LB-LB, EN-LB, DE-LB, or FR-LB instruction tuning data

5 Discussion

We believe that the construction of LUXINSTRUCT represents a significant step forward for resource development in Luxembourgish. Its human-written outputs ensure natural, reliable targets, while the cross-lingual design aids alignment across languages (Section 4).

While its most immediate application lies in enhancing the linguistic accuracy and fluency of models operating in Luxembourgish—through improvements in grammar, orthography, and stylistic coherence—the dataset also serves a broader and arguably more impactful purpose: embedding a culturally grounded and context-aware understanding of Luxembourgish within LLMs.

To this end, Wikipedia functions as a carefully curated repository of both global and local knowledge. The Luxembourgish edition in particular emphasizes topics of national relevance, including local history, governmental institutions, prominent cultural figures, and region-specific traditions.

This is further enriched by the inclusion of news articles, which offer insight into the present-day sociopolitical and cultural landscape of Luxembourg. News sources capture real-time discourse and current events, anchoring language use within a living, evolving context. This allows models trained on the dataset to generate outputs that are timely, accurate, and context-sensitive.

Additionally, the use of dictionary and lexical resources introduces an essential semantic layer. The dictionaries employed provide multilingual translations and disambiguating example sentences for polysemous terms. This enables the model to learn context-dependent meanings more effectively, increasing interpretability and reducing semantic ambiguity.

While the current dataset forms a foundation for instruction tuning in Luxembourgish, future efforts will focus on scaling this work. We aim to apply LUXINSTRUCT at larger scale to existing LLMs, further enriching their Luxembourgish capabilities. Parallel to this, we will continue expanding the dataset in both size and diversity, incorporating new seed sources and adopting emerging data generation techniques.

6 Conclusion

This work presents the development of LUXINSTRUCT, the first cross-lingual instruction tuning dataset tailored for Luxembourgish. By incorporating instructions in English, French and German, the dataset enables cross-lingual model alignment for Luxembourgish. Moreover, we provide empirical evidence demonstrating the advantages of cross-lingual instruction tuning over monolingual approaches in such settings. We hope that both the dataset and our findings serve as a foundation for further advancements in Luxembourgish NLP.

Limitations

The key limitation of our dataset is its restricted diversity, as it currently covers only five task types. This constraint reflects the scarcity of high-quality Luxembourgish resources. We made a conscious decision to prioritize quality—relying on human-generated seed data—over quantity or breadth, avoiding large-scale translation from high-resource languages which often introduces noise or mistranslations.

To add some variation, we included multilingual instructions (in three languages plus Luxembourgish) and varied instruction templates where possible. We view this dataset as a starting point and plan to expand it in future work by incorporating additional tasks and further increasing linguistic and instructional diversity.

Ethical Considerations

Our dataset is constructed from publicly available sources, including news articles and Wikipedia entries, which may contain the names of individuals. We chose not to anonymize this information, as doing so would significantly reduce the contextual richness of the data. Since the content is already accessible in the public domain, we consider its inclusion ethically permissible. Preserving these references is important for maintaining data integrity and ensuring the effectiveness of the dataset for real-world applications

References

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. Association for Computational Linguistics.

Laura Bernardy. 2022. A luxembourgish gpt-2 approach based on transfer learning. Master's thesis, University of Trier, Trier, Germany.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Preprint*, arXiv:2401.07037.

Wikimedia Foundation. Wikimedia downloads.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the relationship between alignment and cross-lingual transfer in multilingual transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. 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, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

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.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schuetze. 2024. LongForm: Effective instruction tuning with reverse instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7056–7078, Miami, Florida, USA. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2024. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Preprint*, arXiv:2409.12958.

Nurkhan Laiyk, Daniil Orel, Rituraj Joshi, Maiya Goloburda, Yuxia Wang, Preslav Nakov, and Fajri Koto. 2025. Instruction tuning on public government and cultural data for low-resource language: a case study in kazakh. *Preprint*, arXiv:2502.13647.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F. Chen. 2025. CrossIn: An efficient instruction tuning approach for cross-lingual knowledge alignment. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 12–23, Abu Dhabi. Association for Computational Linguistics.

Cedric Lothritz and Jordi Cabot. 2025. Testing low-resource language support in llms using language proficiency exams: the case of luxembourgish. *Preprint*, arXiv:2504.01667.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089, Marseille, France. European Language Resources Association.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

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.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025. LuxEmbedder: A cross-lingual approach to enhanced Luxembourgish sentence embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11369–11379, Abu Dhabi, UAE. Association for Computational Linguistics.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for Luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104, Abu Dhabi, UAE. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. 2023. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models . In *2023 IEEE International Conference on Big Data (BigData)*, pages 5684–5693, Los Alamitos, CA, USA. IEEE Computer Society.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20829–20855, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

A LuxInstruct

A.1 Creation Process

Here we provide further details on how LuxInstruct was constructed using 3 different sources: 1) Wikipedia, 2) News Articles, 3) Online Dictionary. The final dataset will be released under a *CC BY-NC 4.0*[5] license.

A.1.1 Wikipedia

For the `Open-Ended` component of LuxInstruct, we use the Luxembourgish subset of the Wikipedia dumps (Foundation)[6], released under the *CC BY-SA 4.0* license[7]. Out of roughly 64,000 articles, we randomly select about 20,000 to generate our samples.

The English instruction generation is carried out using `gpt-4.1-mini`. To ensure reliable extraction of the generated pairs and to avoid formatting inconsistencies, we use function calling with a predefined JSON schema to structure the model's responses in a machine-readable format. The model is guided by the following prompt:

```
Create structured instruction-tuning data for language models. From the text below, extract coherent excerpts that a model might generate in response to a clear, concise instruction. Each excerpt should be a complete, accurate, and natural language response and should be taken directly from the text without altering it. Instructions must be in English, answers in Luxembourgish. Ensure instructions are self-contained and context-independent. The instruction does not need to specify Luxembourgish as the output language.

Input Text:

{text}

Return your findings in JSON format.
```

After generating the English instructions, we apply a series of simple heuristic-based filters to remove low-quality instruction-output pairs. A sample is discarded if it meets any of the following criteria:

- The output is not a string;
- The output contains fewer than 10 words;
- The instruction contains the word `List`;
- The output begins with a lowercase letter;
- The output contains a question mark;
- The output does not end with a full stop;
- The output is not written in Luxembourgish;
- The output is not present in the original Wikipedia article (determined via fuzzy matching, allowing for minor inconsistencies such as punctuation differences or sentence truncation).

[5] https://creativecommons.org/licenses/by-nc/4.0/deed.en

[6] https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.lb

[7] https://creativecommons.org/licenses/by-sa/4.0/

A.1.2 News Articles

We collect articles from RTL[8], a Luxembourgish news platform publishing in Luxembourgish, as well as in French since 2011 and in English since 2018. Since each language has a separate website, articles are not explicitly aligned across languages. To find matching articles, we encode them using OpenAI's `text-embedding-3-small` and select pairs with cosine similarity above 0.65. We discard articles that are too short or too long, and those with titles under six words. The resulting article pairs are used to build the **`Article-To-Title`** and **`Title-To-Article`** datasets in LUXINSTRUCT.

We also create 50 prompt templates in each language (e.g., "`Draft a publication-ready article based on the headline provided:`") and randomly assign one to each pair.

We apply a similar procedure, excluding the cross-lingual article matching, to Luxembourgish-only articles in order to construct the monolingual portions of `article-to-title` and `title-to-article`.

A.1.3 Online Dictionary

The *Luxembourg Online Dictionary* (LOD) provides free online access to Luxembourgish vocabulary, including translations into four languages—French, German, English, and Portuguese—as well as contextual usage examples for numerous terms. The dataset is fully accessible online[9] under the *CC0 1.0* license[10].

A.2 Dataset Statistics

The exact numbers of created samples per instruction language and per task type are provided in Table 1.

A.3 Examples from LUXINSTRUCT

Table 2 contains 2 examples per task type.

B Experiments on Cross-Lingual Alignment

B.1 Models

In our experiments we use the following models:

Qwen3-0.6B[11] (Yang et al., 2025)
A 0.75B-parameter, 28-layer, instruction-tuned model with a 32K context window, trained with 36T tokens and with multilingual support in over 119 languages, including Luxembourgish, released under the *Apache 2.0* license[12].
Gemma-3-1B-IT[13] (Team et al., 2025)
A 1B-parameter, 26-layer, instruction-tuned model with a 128K context window, trained with 2T tokens and with multilingual support in over 140 languages, including Luxembourgish, released with the *Gemma Terms of Use*[14].
OLMo-2-1B-Instruct[15] (OLMo et al., 2025)
A 1.48B-parameter, 16-layer, instruction-tuned model, trained with 4T tokens, with primarily English support, released under the *Apache 2.0* license[16].
Llama-3.2-1B-Instruct[17]
A 1.23B-parameter, 16-layer, instruction-tuned model with a 128K context window, trained with 5T tokens and with multilingual support in 8 languages, released under the *Llama 3.2 Community License*[18].
Phi-4-Mini-Instruct[19] (Microsoft et al., 2025)
A 3.84B-parameter, 32-layer, instruction-tuned model with a 128K context window, trained with 9T tokens and with multilingual support in 23 languages, released under the *MIT License*[20].
Llama-3.1-8B-Instruct[21] (Grattafiori et al., 2024)
A 8.03B-parameter, 32-layer, instruction-tuned model with a 128K context window, trained with 9T tokens and with multilingual support in 8 languages, released under the *MIT License*[22].

B.2 Technical Details

We apply LoRA (Hu et al., 2021) to fine-tune the value, query, and key projections in the attention layers, using a rank of 8, scaling factor $\alpha = 16$, and a dropout rate of 0.05. Each model is trained

[8] https://www.rtl.lu
[9] https://data.public.lu/en/datasets/letzebuerger-online-dictionnaire-lod-linguistesch-daten/
[10] https://creativecommons.org/publicdomain/zero/1.0/
[11] https://huggingface.co/Qwen/Qwen3-0.6B
[12] https://www.apache.org/licenses/LICENSE-2.0
[13] https://huggingface.co/google/gemma-3-1b-it
[14] https://ai.google.dev/gemma/terms
[15] https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct
[16] https://www.apache.org/licenses/LICENSE-2.0
[17] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[18] https://www.llama.com/llama3_2/license/
[19] https://huggingface.co/microsoft/Phi-4-mini-instruct
[20] https://opensource.org/license/mit
[21] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[22] https://www.llama.com/llama3_1/license/

	de	fr	en	lb	Total
`Article-To-Title`		15 468	6 826	48 993	71 287
`Title-To-Article`	19 107	9 090	49 002		77 199
`Colloquial-To-Standard`	2 318	2 318	2 318	2 318	9 272
`Word-To-Example`	40 489	40 454	38 529	40 528	160 000
`Open-Ended`	20 723	20 723	58 898	20 723	121 067
Total	63 530	98 070	115 661	161 564	438 825

Table 1: Data distribution across different languages and task types

for 500 steps with a batch size of 16, a learning rate of $2e^{-5}$, weight decay of 0.01, and a context length of 128 tokens.

To compute cross-lingual alignment scores between language pairs, we use the *devtest* split of the FLORES-200 dataset[23] (Team et al., 2022), which contains 1 012 parallel sentences across 204 languages, including Luxembourgish. Document-level representations are obtained by mean-pooling the final-layer contextualized token embeddings. Alignment between languages is quantified using the Centered Kernel Alignment (CKA) metric (Kornblith et al., 2019).

All experiments are conducted on a single Nvidia T4 GPU and complete within a few hours.

B.3 Full Results

In Table 3 we provide the full results that have been summarized in Figure 2.

[23] https://huggingface.co/datasets/facebook/flores

Task Type	Example	
	Instruction	**Output**
`Open-Ended`	Since when has Murcia been an autonomous community?	Murcia ass zanter dem 9. Juni 1982 eng autonom Gemeinschaft.
	Who was Thomas Keith Glennan and what was his role at NASA?	Den Thomas Keith Glennan, gebuer den 8. September 1905 zu Enderlin am North Dakota, USA, a gestuerwen den 11. Abrëll 1995 zu Mitchellville, Maryland, war tëscht August 1958 a Januar 1961 den éischten Administrateur vun der US-Raumfaartagentur NASA.
`Article-To-Title`	Turn the following news article into a concise headline: "Greece will have to be patient before it gets the next installment of its European financial aid. Following a meeting of the EU ministers..."	Griicheland muss sech nach e bësse gedëllegen
	Pick a headline that would summarize the article below: "Several commemoration ceremonies took place last Sunday, in order to celebrate Luxembourg's national resistance day. The commemoration ceremony doesn't vary much over the years. However, recent images..."	Erënnerungen u Krich an Ënnerdréckung héichhalen
`Title-To-Article`	Use this news headline to inspire a detailed article: "Social media to comply with new EU regulations"	Video-Plattforme wéi Youtube musse sech an der EU an Zukunft u méi strikt Reegele beim Jugendschutz oder och bei Reklammen halen. Déi zoustänneg Kommissioun vum Europaparlament...
	Produce an informative and factual story using this title: "Police looking for driver involved in pedestrian hit and run"	Zu Wolz gouf eng Persoun op engem Zebrasträife ugestouss. Ouni sech ëm d'Affer ze këmmeren, ass den Auto einfach fortgefuer. En Donneschdeg de Moien um kuerz virun 11 Auer...
`Word-To-Example`	Demonstrate usage of the Luxembourgish word "eethesch" (translation: "ethical") in a sentence.	Den Asaz vu Kënschtlecher Intelligenz bréngt dacks eethesch a sozial Erausfuerderunge mat sech.
	Use the term "Fuerscherin" in a Luxembourgish sentence, translating to "researcher".	Déi jonk Fuerscherin sicht mat hirem Team no Léisunge géint de Klimawandel.
`Colloquial-To-Simplified`	Clarify the meaning of this informal sentence: "Him ass eng gutt Geleeënheet laanscht d'Nues gaangen."	Hien huet eng gutt Geleeënheet verpasst.
	Rephrase the following colloquial sentence to make it easier to understand: "Den Informatiker huet de Computer mat Date gefiddert."	Den Informatiker huet Daten an de Computer aginn.

Table 2: Examples from LUXINSTRUCT for each task type

Model	Training Data	Compared embedding spaces		
		LB-DE	**LB-EN**	**LB-FR**
Qwen 3 (0.6B)	Base	0.2612	0.2342	0.2198
	DE-LB	0.3542	0.3456	0.3257
	EN-LB	**0.3909**	**0.3610**	**0.3378**
	FR-LB	0.3894	0.3587	0.3033
	LB-LB	0.3822	0.3528	0.3347
Llama 3.2 (1B)	Base	0.2359	0.2155	0.2091
	DE-LB	0.2649	0.2912	0.2754
	EN-LB	**0.3332**	**0.3222**	**0.3076**
	FR-LB	0.3302	0.3197	0.2871
	LB-LB	0.2950	0.2759	0.2678
Gemma 3 (1B)	Base	0.2774	0.2303	0.2144
	DE-LB	0.2981	0.2488	0.2255
	EN-LB	0.3029	**0.2570**	0.2264
	FR-LB	**0.3035**	0.2555	**0.2290**
	LB-LB	0.3027	0.2490	0.2292
OLMo 2 (1B)	Base	0.3020	0.2666	0.2784
	DE-LB	0.3381	0.3544	0.3429
	EN-LB	0.3639	0.3555	**0.3438**
	FR-LB	**0.3682**	**0.3665**	0.3003
	LB-LB	0.3582	0.3384	0.3376
Phi-4-mini (1.8B)	Base	0.2090	0.1787	0.1879
	DE-LB	0.3494	0.3262	0.3196
	EN-LB	0.3600	0.3345	0.3273
	FR-LB	**0.3815**	**0.3547**	**0.3466**
	LB-LB	0.2798	0.2541	0.2539
Llama 3.1 (8B)	Base	0.2620	0.2278	0.2381
	DE-LB	0.4046	0.4177	0.4292
	EN-LB	**0.4747**	**0.4315**	**0.4433**
	FR-LB	0.4496	0.4075	0.3709
	LB-LB	0.4520	0.3975	0.4113

Table 3: CKA values for various models and training data configurations. Bold values indicate the highest per column within each model.