LUXINSTRUCT: A Cross-Lingual Instruction Tuning Dataset For Luxembourgish

Fred Philippy[1] **Laura Bernardy**[1] **Siwen Guo**[2]

Jacques Klein[1] **Tegawendé F. Bissyandé**[1]

[1]SnT, University of Luxembourg, Luxembourg

[2]Luxembourg Institute of Science and Technology, Luxembourg

Correspondence: fred.philippy@uni.lu

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 resorting to machine-generated translations into it. 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 cross-lingual instruction tuning not only improves representational alignment across languages but also the model's generative capabilities in Luxembourgish. This highlights how cross-lingual data curation can avoid the common pitfalls of machine-translated data and directly benefit low-resource language development.[1]

1 Introduction

Instruction tuning improves the ability of Large Language Models (LLMs) to follow user prompts by fine-tuning them on curated instruction–output pairs, leading to better generalization and alignment 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 such resources are abundant, low-resource languages have very few. 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 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.

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

To address the scarcity of high-quality data, we compile LUXINSTRUCT, a cross-lingual instruction tuning dataset for Luxembourgish, in which instructions are written in other languages and outputs are in Luxembourgish. By avoiding machine translation into Luxembourgish[2], 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 improved model performance compared to monolingual MT-based instruction tuning data.

[1]Code and data are provided at https://github.com/fredxlpy/LuxInstruct.

[2]MT into Luxembourgish is limited to an ablation subset of instructions and omitted from LUXINSTRUCT.

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 et al. (2026) 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

Although 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 MT (Li et al., 2023; Holmström and Doostmohammadi, 2023; Li et al., 2024) or repurpose labeled NLP datasets (Muennighoff et al., 2023). However, neither method is effective for Luxembourgish, due to limited MT 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 the *Multilingual Reverse Instructions* (MURI) framework (Köksal et al., 2025). Yet, MURI still relies on two rounds of MT 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., 2025; Lin et al., 2025).

3 LuxInstruct

3.1 Dataset Creation

We create LuxInstruct using three primary Luxembourgish data sources: Wikipedia, News Articles, and an Online Dictionary. More information on the source data and the process is provided in Appendix A.

Wikipedia We adopt a reverse instruction generation approach inspired by the MURI framework (Köksal et al., 2025), but diverge in key aspects to avoid translation artifacts. Instead of first translating the source data into English and then translating the generated instructions into the target language, as in MURI, we simply prompt OpenAI's `gpt-4.1-mini`[3] to extract informative content directly from Luxembourgish Wikipedia and generate corresponding instructions in English. This allows for high-quality semantically aligned instruction-output pairs without relying on MT. 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. Then, we additionally machine-translate a subset of the instructions to German and French using `gpt-4.1-mini`. The resulting dataset forms the **`Open-Ended`** portion of LuxInstruct.

News Articles The Luxembourgish news platform *RTL Luxembourg* 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 article pairs, we 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`**). In addition, similar monolingual Luxembourgish instruction-output pairs are created.

Furthermore, we leverage the LuxAlign (Philippy et al., 2025) corpus of parallel

[3] In preliminary tests, we found that `gpt-4.1-mini` cannot reliably generate Luxembourgish text, but reliably understands it well enough for this extraction task.

Luxembourgish-French-English semantically similar sentences derived from news articles to create a cross-lingual paraphrasing task (**`CL-Paraphrase`**).

Online Dictionary We leverage a publicly available Luxembourgish dictionary containing lexical entries with synonyms, translations (to English, French, German) and example sentences. We design three 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`**), (2) simplifying colloquial Luxembourgish sentences (**`Colloquial-To-Standard`**), and (3) word translation to Luxembourgish (**`Word-Translation`**).

3.2 Dataset Statistics

Our new dataset consists of 391,551 cross-lingual instruction-output samples across English, French, and German as instruction languages, along with 145,793 monolingual samples in Luxembourgish, where both instruction and output are in Luxembourgish. Appendix A provides the exact number of samples per language and type of task (Table 2) as well as examples (Table 4). LuxInstruct content derived from Wikipedia and the Online Dictionary is licensed under *CC BY-SA 4.0*, whereas the News Articles content is non-commercial.

4 Cross-Lingual Vs Monolingual Instruction Tuning

We compare cross-lingual and monolingual instruction tuning for Luxembourgish through two experiments: one measuring cross-lingual alignment (§4.1) and one evaluating instruction-following in in-context learning (§4.2). Both experiments use the `Open-Ended` portion, containing parallel English, French, and German instructions, further extended to Luxembourgish via `gpt-4.5`.

4.1 Cross-Lingual Alignment

Each model is fine-tuned on same-sized subsets, with instructions in a single language (EN, FR, DE, or LB) and responses 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 (NLLB 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 are provided in Appendix B.1.

Figure 1 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 5 in the Appendix.

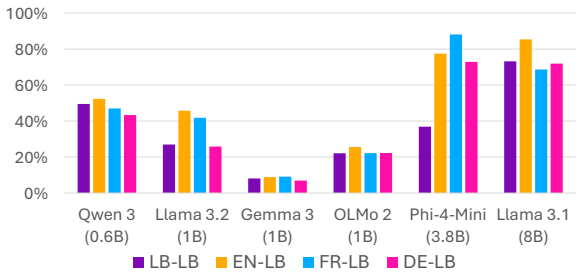

Figure 1: Mean variation (in %) in alignment between the Luxembourgish and the English, French, and German representation spaces after fine-tuning on LB-LB, EN-LB, DE-LB, or FR-LB instruction tuning data

4.2 Few-Shot In-Context Learning

To evaluate the impact of instruction language in instruction tuning on an LLM's text generation capabilities, we adopt a few-shot in-context learning strategy. More specifically, we compare different instruction languages (`LB`, `EN`, `FR`, `DE`, and a mixed-language setting `MULTI`) for these few-shot examples, while keeping the output language fixed to Luxembourgish.

We use `Gemma 3 12B` & `27B` (Gemma Team et al., 2025) for these experiments, as they are the only moderately sized open-source models among those tested that yield reasonable generation quality in Luxembourgish.

Given the absence of a standard instruction-following benchmark in Luxembourgish, we construct a set of 50 manually verified instructions. Since standard metrics are inapplicable, we employ the G-Eval framework (Liu et al., 2023), with evaluations conducted via `GPT-5 mini`, `Gemini 2.5 Flash-Lite` (Comanici et al., 2025), and `DeepSeek-V3` (DeepSeek-AI et al., 2025). Using

three LLM judges improves robustness, particularly in low-resource settings where evaluation may be less reliable than in English. To further validate this approach, Appendix B.2.5 presents a human annotation study that verifies the consistency between human assessments and the aggregated judgments of the three LLM evaluators.

Table 1 presents the average scores from all three evaluators across four different dimensions. Full model-specific scores are available in Table 6 in the Appendix. To enhance result robustness, we repeat the experiments five times with different seeds (and corresponding few-shot examples), reporting the mean scores across all runs.

The findings clearly show that cross-lingual few-shot examples outperform monolingual Luxembourgish ones, demonstrating not only the utility of LUXINSTRUCT, but also reinforcing our decision to pursue a cross-lingual strategy.

	k	IT Lang.	CL	CO	FL	RE
Gemma 12B	0-Shot		60.1	67.1	57.3	69.2
	4	LB	61.6	70.9	59.4	74.0
		DE	61.6	72.2	60.4	75.5
		EN	**63.4**	**73.2**	**61.8**	76.0
		FR	62.0	72.6	60.7	**76.1**
		MULTI	61.6	71.8	59.4	74.5
	8	LB	62.1	72.1	60.3	75.7
		DE	61.7	72.7	60.8	76.3
		EN	**64.7**	**74.8**	**63.0**	**77.9**
		FR	62.6	73.2	61.9	76.9
		MULTI	62.4	72.9	61.2	76.5
Gemma 27B	0		68.9	74.2	67.1	75.1
	4	LB	77.6	83.5	75.6	85.2
		DE	76.6	83.0	74.7	85.0
		EN	**78.7**	**85.1**	**77.5**	**86.8**
		FR	77.1	84.3	74.8	85.3
		MULTI	78.1	84.0	75.7	85.5
	8	LB	78.9	85.8	76.4	87.1
		DE	80.5	87.8	78.4	**89.6**
		EN	**81.3**	**88.7**	**79.9**	89.3
		FR	78.1	85.6	76.8	87.3
		MULTI	80.1	86.4	78.0	88.2

Table 1: G-Eval results measuring instruction-following performance in Luxembourgish across Clarity (**CL**), Coherence (**CO**), Fluency (**FL**), and Relevance (**RE**) in 0-shot, 4-shot, and 8-shot settings, with few-shot examples using different instruction languages (**IT Lang.**).

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).

Its most immediate application is improving the linguistic accuracy and fluency of models in Luxembourgish, particularly in grammar, orthography, and stylistic consistency. At the same time, the dataset plays a broader role by embedding a culturally grounded and context-aware understanding of the language within LLMs.

To this end, Wikipedia serves as a curated source of both global and local knowledge. The Luxembourgish edition emphasizes nationally relevant topics, including local history, government institutions, prominent cultural figures, and regional traditions.

This foundation is further strengthened by news articles, which provide insight into the current sociopolitical and cultural landscape of Luxembourg. News sources reflect real-time discourse and events, anchoring language use in a dynamic, evolving context. This allows models to produce outputs that are timely, accurate, and contextually informed.

Additionally, lexical and dictionary resources introduce an essential semantic layer. The dictionaries used include multilingual translations and clarifying examples for polysemous terms, helping the model capture context-specific meanings and reduce 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 on a 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 seven 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, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2025. XCOT: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23550–23558.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 technical report. *Preprint*, arXiv:2412.19437.

Wikimedia Foundation. Wikimedia Downloads.

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.

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.

Jeffrey Ip and Kritin Vongthongsri. 2025. DeepEval. The Open-Source LLM Evaluation Framework.

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.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. MURI: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14509–14538, Vienna, Austria. Association for Computational Linguistics.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in LLMs using language proficiency exams: the case of Luxembourgish. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2453–2476, Rabat, Morocco. Association for Computational Linguistics.

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, Xiyang Dai, and 56 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

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.

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. These sources contain high-quality Luxembourgish text and are inherently safety-filtered, removing the need for an additional safety filtering step.

A.1.1 Wikipedia

For the `Open-Ended` component of LuxInstruct, we use the Luxembourgish subset of the Wikipedia dumps (Foundation)[4], released under the *CC BY-SA 4.0* license[5]. 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`, which was selected due to its computational efficiency and cost-effectiveness. In a small manual qualitative preliminary evaluation, we found that it demonstrates sufficiently strong Luxembourgish comprehension. As the task requires only understanding, not generating, Luxembourgish text, its generative capabilities in Luxembourgish are not a priority. 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:

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

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

```
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).

While we use a language identification model to ensure outputs are predominantly Luxembourgish, natural code-switching is intentionally preserved, as it is inherent to authentic Luxembourgish usage.

A.1.2 News Articles

We collect articles from RTL[6], 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`**.

The **`CL-Paraphrase`** component leverages the LUXALIGN dataset (Philippy et al., 2025), which includes 28.2k Luxembourgish-English and 89.4k Luxembourgish-French semantically aligned sentence pairs. Here too, we generate 50 prompt templates (e.g., `"Modify the sentence with new wording:"`) and assign them randomly to each bilingual pair, where the input is either English or French and the output is Luxembourgish.

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[7] under the *CC0 1.0* license[8].

All LUXINSTRUCT components sourced from LOD, namely **`Colloquial-To-Standard`**, Word-To-Example, and **`Word-Translation`**, are generated using 20 prompt templates per language for the first task and 50 prompt templates per language for the latter two tasks.

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

[7] https://data.public.lu/en/datasets/letzebuerger-online-dictionnaire-lod-linguistesch-daten/

[8] https://creativecommons.org/publicdomain/zero/1.0/

A.2 Dataset Statistics

The exact numbers of created samples per instruction language and per task type are provided in Table 2.

A.3 Examples from LUXINSTRUCT

Table 4 contains 2 examples per task type.

B Experimental Setup

We conduct our experiments on a subset of the `Open-Ended` section of LUXINSTRUCT, restricted to the samples where instructions have additionally been translated into French and German. For the purposes of these experiments, we additionally translate these instructions into Luxembourgish using `gpt-4.5`, though these translations are not included in the final dataset.

B.1 Cross-Lingual Alignment

B.1.1 Models

In our experiments we use the following models:

Qwen3-0.6B[9] (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[10].
Gemma-3-1B-IT[11] (Gemma Team et al., 2025)
A 1B-parameter, 26-layer, instruction-tuned model with a 32K context window, trained with 2T tokens and with multilingual support in over 140 languages, including Luxembourgish, released with the *Gemma Terms of Use*[12].
OLMo-2-1B-Instruct[13] (Team 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[14].
Llama-3.2-1B-Instruct[15]
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*[16].
Phi-4-Mini-Instruct[17] (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*[18].
Llama-3.1-8B-Instruct[19] (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*[20].

B.1.2 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 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[21] (NLLB 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.1.3 Full Results

In Table 5 we provide the full results that have been summarized in Figure 1.

B.2 Few-Shot In-Context Learning

B.2.1 Text Generation Model

In this experiment we use **`Gemma 3`** (Gemma Team et al., 2025) in its 12B and 27B versions, with 48 and 62 layers and trained on 12T and 14T tokens,

[9] https://huggingface.co/Qwen/Qwen3-0.6B
[10] https://www.apache.org/licenses/LICENSE-2.0
[11] https://huggingface.co/google/gemma-3-1b-it
[12] https://ai.google.dev/gemma/terms
[13] https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct
[14] https://www.apache.org/licenses/LICENSE-2.0
[15] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct
[16] https://www.llama.com/llama3_2/license/
[17] https://huggingface.co/microsoft/Phi-4-mini-instruct
[18] https://opensource.org/license/mit
[19] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[20] https://www.llama.com/llama3_1/license/
[21] https://huggingface.co/datasets/facebook/flores

	de	fr	en	lb	Total
`Article-To-Title`		11 774	4 744	51 498	68 016
`Title-To-Article`		12 891	4 899	51 507	69 297
`Colloquial-To-Standard`	2 323	2 323	2 323	2 323	9 292
`Word-To-Example`	40 426	40 390	38 418	40 465	159 699
`Open-Ended`	19 955	19 955	56 633		96 543
`Word-Translation`	6 153	5 840	4 927		16 920
`CL-Paraphrase`		89 405	28 172		117 577
Total	68 857	182 578	140 116	145 793	537 344

Table 2: Data distribution across different languages and task types

respectively. Both models are instruction-tuned with a 128K context window, support over 140 languages (including Luxembourgish), and include visual capabilities.

B.2.2 Test Data

Due to the absence of high-quality instruction-following benchmarks in Luxembourgish, a native speaker was engaged to curate a set of 50 Luxembourgish test instructions, which were subsequently used to evaluate the model's performance.[22]

B.2.3 Technical Details

To accommodate resource limitations, the model used for text generation is quantized to 4-bit precision. We perform generation using parameters $top_k = 64$, $top_p = 0.95$, and temperature $= 1.0$, with a maximum output length of 500 tokens.

Since we machine-translated a subset of the instructions from the `Open-Ended` section of LUX-INSTRUCT, we are able to use the same few-shot in-context examples across all instruction language settings.

This design choice enables us to isolate the influence of instruction language while controlling for few-shot examples selection variability. To further enhance the reliability of our results, we perform five independent generations using five distinct sets of few-shot examples, and incorporate all generated outputs in the subsequent evaluation.

In the `LB`, `FR`, `EN`, and `DE` settings, all few-shot instructions are provided solely in the respective languages, with outputs always in Luxembourgish. In contrast, the `MULTI` setting combines instructions across languages: one per language in the 4-shot setup and two per language in the 8-shot setup.

[22]The exact test instructions are provided here.

B.2.4 G-Eval

For evaluation, we adopt the G-Eval framework (Liu et al., 2023), which employs an LLM-as-a-judge approach. We use its implementation via the `DeepEval` library (Ip and Vongthongsri, 2025), using it to assess outputs along four dimensions: Clarity, Relevance, Fluency, and Coherence. Scores are returned on a 0–1 scale, which we multiply by 100 for reporting. To enhance robustness, we conduct evaluations using three different LLM APIs: `GPT-5 mini`, `Gemini 2.5 Flash-Lite` (Comanici et al., 2025), and `DeepSeek-V3` (DeepSeek-AI et al., 2025).

The specific Chain-of-Thought prompts used across the four evaluation dimensions are as follows:

Clarity

- `Compare the language used in the output with the input to assess clarity and simplicity.`
- `Check if the output presents information logically and in an organized manner relative to the input.`
- `Identify any ambiguous or unclear phrases in the output that deviate from the meaning of the input.`
- `Ensure that the formatting and structure of the output enhance understanding compared to the input.`

Relevance

- Compare the output to the input instruction to determine if the output addresses the task requested.
- Verify that the content of the output is contextually relevant and directly related to the input instruction.
- Check for completeness by ensuring the output fully responds to the elements specified in the input instruction.
- Assess clarity and coherence of the output in relation to the input instruction, confirming it makes sense as a response.

Fluency

- Compare the output to the input to ensure the content is relevant and appropriately reflects the input context.
- Check the grammar in the output for correctness including verb tense, subject-verb agreement, and sentence structure.
- Evaluate the coherence of the output by assessing logical flow and clarity throughout the response.
- Determine if the output reads naturally and smoothly as a fluent piece of text in relation to the input.

Coherence

- Compare the input with the output to determine if the output logically addresses the content and intent of the input.
- Assess the structure of the output for clear organization and progression of ideas relative to the input.
- Evaluate whether the output maintains a consistent and coherent flow that aligns with the input's context and purpose.
- Identify any gaps or contradictions between the input and output that disrupt logical coherence.

We provide the full evaluation results in Table 6.

B.2.5 Human Calibration of G-Eval

To assess the reliability of LLM-based judgments, we complemented our automatic evaluation with a small-scale human study. Our working assumption is that the consensus formed when all three judge LLMs prefer the same outputs offers a more stable and trustworthy indicator of quality than using only one model. To evaluate this assumption, we randomly sampled output pairs on which all judge LLMs agreed about the preferred candidate (i.e., one output consistently received higher scores across dimensions). Human annotators were then asked to assess the same pairs, without any knowledge of the LLMs' choice. To reduce annotation effort, we merged highly correlated evaluation dimensions: clarity and fluency were combined into one metric, while relevance and coherence formed another. For each pair, the annotator chose between three options: preference for the first output, preference for the second, or a tie. We then measured agreement between the human annotations and the LLM judges using Cohen's κ. Since the LLMs do only rarely produce ties (due to the continuous nature of their scores), tie cases from the human annotations were excluded from the κ calculation.

Human Annotator	CL & FL	RE & CO
A	0.957	0.935
B	0.863	0.780

Table 3: Cohen's κ scores measuring agreement between two human annotators and the LLM judges for two sets of dimensions: Clarity and Fluency (CL & FL), and Relevance and Coherence (RE & CO)

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-Standard`	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.
`CL-Paraphrase`	Rework this sentence in a new form: "Luxembourg's population increased by 10,667 people in 2021."	D'Awunnerzuel zu Lëtzebuerg ass 2021 ëm 10.667 Persounen eropgaangen.
	Paraphrase the following: "He had been in shock himself, under a lot of stress and scared."	Hie selwer hätt ënner Schock, Stress an Angscht gestanen.
`Word-Translation`	Find 2 Luxembourgish translations for the English word "appreciation [judgement, evaluation]".	"Aschätzung" an "Appreciatioun"
	List 3 ways "consent" can be said in Luxembourgish.	"Autorisatioun", "Geneemegung" an "Awëllegung"

Table 4: 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 5: CKA values for various models and training data configurations. Bold values indicate the highest per column within each model.

Judge	Shots	IT Lang.	Gemma 3 12B				Gemma 3 27B			
			CL	CO	FL	RE	CL	CO	FL	RE
DeepSeek-V3	0-Shot		78.04	80.20	77.64	81.96	76.64	78.56	79.40	76.56
	4-Shot	LB	81.16	85.36	81.00	87.40	88.36	89.40	91.48	90.52
		DE	81.84	86.92	82.28	**89.72**	88.24	87.96	91.32	90.68
		EN	82.32	86.88	**82.92**	89.20	**88.92**	90.40	**92.76**	**91.56**
		FR	**82.36**	**86.96**	82.88	89.44	88.44	**91.28**	91.16	90.12
		MULTI	82.04	85.96	81.84	88.68	88.28	89.40	90.88	89.92
	8-Shot	LB	82.32	86.88	82.28	89.40	91.16	92.16	92.84	93.96
		DE	81.96	86.68	83.16	90.20	92.80	94.68	95.24	**96.16**
		EN	**83.64**	**88.40**	**84.64**	**91.00**	**93.76**	**94.88**	**95.52**	95.40
		FR	82.24	86.88	83.64	90.32	90.56	92.48	93.88	93.92
		MULTI	83.04	87.32	83.68	90.36	91.68	92.04	93.44	93.96
Gemini 2.5 Flash-Lite	0-Shot		62.92	71.32	61.32	74.48	74.04	78.88	73.04	82.92
	4-Shot	LB	65.16	77.52	64.16	83.08	82.68	89.48	81.24	91.64
		DE	65.20	79.16	65.28	82.88	80.28	88.32	79.56	90.44
		EN	**67.36**	**79.76**	**67.40**	83.80	**83.68**	**90.80**	**83.60**	**93.48**
		FR	64.44	78.40	64.36	**83.88**	81.44	89.88	80.76	93.16
		MULTI	64.26	78.80	63.96	82.76	82.72	90.32	82.28	92.76
	8-Shot	LB	65.18	79.60	65.74	83.84	83.52	91.44	82.28	93.32
		DE	64.56	80.36	65.28	84.60	84.16	93.44	85.04	**95.76**
		EN	**68.52**	**81.64**	**69.20**	**86.36**	**84.52**	**94.10**	**85.92**	94.88
		FR	65.90	80.40	67.68	85.12	81.16	91.04	81.60	93.36
		MULTI	65.00	79.20	65.48	84.44	83.32	92.08	84.20	93.56
GPT-5 mini	0-Shot		39.32	49.88	33.04	51.04	55.96	65.24	48.80	65.72
	4-Shot	LB	38.56	49.92	32.96	51.56	61.84	71.76	54.20	73.48
		DE	37.76	50.52	33.60	53.96	61.28	72.80	53.16	73.76
		EN	**40.60**	**53.04**	**35.16**	54.88	**63.60**	**74.20**	**56.00**	**75.28**
		FR	39.16	52.52	34.80	**55.12**	61.36	71.84	52.52	72.72
		MULTI	38.56	50.56	32.36	52.00	63.20	72.24	53.92	73.88
	8-Shot	LB	38.72	49.88	33.00	53.88	61.96	73.76	54.08	74.00
		DE	38.64	51.16	33.88	54.16	64.64	75.36	55.04	76.84
		EN	**41.96**	**54.28**	**35.24**	**56.20**	**65.48**	**77.20**	**58.36**	**77.56**
		FR	39.72	52.36	34.48	55.24	62.68	73.36	54.84	74.68
		MULTI	39.24	52.12	34.56	54.68	65.20	75.20	56.48	77.16

Table 6: G-Eval results measuring instruction-following performance across Clarity (**CL**), Coherence (**CO**), Fluency (**FL**), and Relevance (**RE**) in 0-shot, 4-shot, and 8-shot settings, with few-shot examples using different instruction languages (**IT Lang.**).