LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data

Julian Valline, Cedric Lothritz, Siwen Guo, Jordi Cabot
Luxembourg Institute of Science and Technology
5 Av. des Hauts-Fourneaux
L-4362 Esch-sur-Alzette
{julian.valline, cedric.lothritz,
siwen.guo, jordi.cabot}@list.lu

arXiv:2510.24434v4 [cs.CL] 22 Aug 2026

Abstract

The effectiveness of instruction-tuned Large Language Models (LLMs) is often limited in low-resource linguistic settings due to a lack of high-quality training data. We introduce LuxIT, a monolingual instruction tuning dataset for Luxembourgish developed to mitigate this challenge. We synthesize the dataset from a corpus of native Luxembourgish texts, utilizing DeepSeek-R1-0528, chosen for its shown proficiency in Luxembourgish. Following generation, we apply a quality assurance process, employing an LLM-as-a-judge approach, retaining 227,507 high-quality instruction-answer pairs. To investigate the practical utility of the dataset, we fine-tune 14 smaller-scale LLMs ($\leq$15B parameters) on LuxIT and evaluate them on standardized Luxembourgish proficiency exams and five downstream NLP tasks. Training on LuxIT yields a mean accuracy change of +5.37 percentage points on language exams across all 14 models, with 12 of 14 showing improvement. On NLP downstream tasks, 9 of 14 models improve in macro-averaged F1, though gains on the two benchmarks do not systematically correlate. These results underscore the feasibility of leveraging monolingual synthetic data to improve LLM capabilities in low-resource languages, while highlighting the multi-faceted nature of language proficiency.

1 Introduction

In recent years, Large Language Models (LLMs) have demonstrated remarkable proficiency across a diverse range of natural language tasks (Zhang et al., 2025; Brown et al., 2020; Touvron et al., 2023b; Yin et al., 2023; Peng et al., 2023; Li et al., 2023a). Their rapid and widespread adoption as virtual assistants (Shu et al., 2023) has been largely driven by their accessibility through intuitive chat-based interfaces (Weber et al., 2024; Touvron et al., 2023b). The training of these assistants involves multiple steps, beginning with a pre-training stage where the model learns from a vast text corpus using a self-supervised objective such as next-word prediction (Weber et al., 2024; Touvron et al., 2023b; Zhou et al., 2023). The subsequent, critical step is instruction tuning (IT) (Wei et al., 2022) which refines the model's ability to follow user instructions and engage in a conversational manner (Weber et al., 2024; Zhang et al., 2025). This process involves fine-tuning an LLM on an extensive dataset, which typically consists of instruction-answer pairs. By training on these pairs, the model learns to generate appropriate responses to user instructions, resulting in more anticipated and manageable outputs (Zhang et al., 2025). A notable benefit of this approach is that it can enhance the model's ability to generalize and respond to previously unseen instructions (Nayak et al., 2024; Zhou et al., 2023; Sanh et al., 2022).

A significant issue at present is the scarcity of open-source, instruction tuning datasets for underrepresented languages. Existing datasets are predominantly in English, marginalizing other languages and leading to inferior model performance and higher deployment costs in these settings (Weber et al., 2024). In the context of this work, "low-resource" refers to a scarcity of NLP tooling, open-source datasets, and specialized instruction-tuning resources. Translating an existing English instruction set, or pairing target-language content with a higher-resource language both depend on resources that are themselves scarce in these settings (Köksal et al., 2025). Luxembourgish, a West Germanic language spoken by about 600,000 people[1] primarily in Luxembourg, exemplifies this challenge as adequate training data is scarce.

In this paper, we introduce LuxIT, a monolingual instruction tuning dataset in Luxembourgish,

[1] https://cursus.edu/en/23040/luxembourgish-at-its-best, as of 11.08.2026

synthesized from articles of the news website RTL[2] and Wikipedia entries. Following the work of Lothritz et al. (2026), we select DeepSeek-R1-0528, as the top-performing model for Luxembourgish generation and comprehension to synthetically produce a final dataset of 227,507 high-quality instruction-answer pairs. To validate the quality of our dataset, we assess the Luxembourgish capabilities of several LLMs before and after they have been fine-tuned on LuxIT.

To our knowledge, LuxIT is the first Luxembourgish instruction tuning dataset generated entirely within the target language, requiring neither a machine translation system nor parallel data.

Our primary contributions are twofold:

- We introduce LuxIT[3], a synthetically generated instruction tuning dataset in Luxembourgish.
- We evaluate the utility of LuxIT by fine-tuning 14 widely-used small-scale LLMs ($\leq$ 15B parameters) on LuxIT and evaluating them on standardized Luxembourgish proficiency exams and five downstream NLP tasks, demonstrating improvements in language proficiency for 12 of 14 models and NLP task performance for 9 of 14 models.

To validate the pipeline's ability to produce high-quality data, we assess the generated outputs against four custom metrics (see Section 3.4), quantifying the proportion of retained samples. We then address the following research questions:

RQ1: How does fine-tuning on LuxIT affect performance on Luxembourgish language exams?

i) Does fine-tuning on LuxIT improve overall accuracy across CEFR proficiency levels? We fine-tune 14 LLMs ($\leq$ 15B parameters) on LuxIT and benchmark their performance in Luxembourgish against their corresponding base instruct models, following the evaluation framework by Lothritz et al. (2026).

ii) How does the impact vary across individual linguistic categories? We examine the results from RQ1 further and analyze them over the individual categorical exam levels, akin to Lothritz et al. (2026).

RQ2: Does fine-tuning on LuxIT improve performance on downstream Luxembourgish NLP tasks? We benchmark our fine-tuned models against their respective base instruct models on 5 well-known NLP tasks in Luxembourgish.

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

[3] We provide the Wikipedia subset, RTL subset and code

2 Related work

2.1 Multilingual instruction tuning datasets

xP3 (Muennighoff et al., 2023) is a multilingual, human-crafted instruction tuning dataset, combining data from existing datasets, such as P3 (Sanh et al., 2022) into a unified format across 46 languages. The same work additionally introduces xP3mt, by applying machine translation to the English prompt templates. Bactrian-X (Li et al., 2023b) machine-translates English instructions from Alpaca (Taori et al., 2023) and Dolly (Conover et al., 2023) into 51 other languages. Responses are then generated by GPT-3.5-Turbo. LIMA (Zhou et al., 2023) is a fine-tuned version of Llama (Touvron et al., 2023a), trained on 1000 attentively selected instruction-answer pairs. Weber et al. (2024) extend this resource to Lima-X by translating the LIMA instructions into 4 languages. Comparative overviews of these and related resources are provided by Zhang et al. (2025) and Weber et al. (2024). However, none of these resources covers Luxembourgish.

2.2 Luxembourgish language resources

Recently, there have been various contributions to the creation of Luxembourgish language resources. Lothritz et al. (2022) developed a pretraining dataset in Luxembourgish through a data-augmentation technique where they partially translated text from German into Luxembourgish, resulting in LuxemBERT. Plum et al. (2025) presented LuxGen, a benchmark for evaluating data generation in Luxembourgish and LuxT5, an mT5-based (Xue et al., 2021) text generation model in Luxembourgish, pre-trained on a German, French and Luxembourgish text corpus, where the latter is obtained through transfer learning from German and French. Most recently, Philippy et al. (2025b) introduced a hand-crafted cross-lingual dataset for training Luxembourgish sentence embedding models, which led to the LuxEmbedder model.

With the increasing ubiquity of LLMs, there is a need for a high-quality instruction tuning dataset in Luxembourgish to guide weaker models in improving their Luxembourgish capabilities. Two recent approaches generate Luxembourgish instruction data from native texts. Köksal et al. (2025) ap-

ply reverse instructions to entries from Wikipedia and CulturaX (Nguyen et al., 2024) across 200 languages, including Luxembourgish. Their pipeline however relies on machine translating the article into English and the generated instruction back into the target language. The cross-lingual dataset from Philippy et al. (2025a) demonstrates an effective strategy for generating instruction data in Luxembourgish without machine translation, pairing Luxembourgish with English, German, or French content. Yet, this approach is contingent on the availability of parallel seed data, which is often a bottleneck for low-resource languages. Our work complements both approaches by performing generation directly in Luxembourgish, requiring neither machine translation nor parallel data.

3 LuxIT

This section details the methodology employed for the creation of our Luxembourgish instruction tuning dataset. The data generation process for LuxIT, illustrated in Figure 1 is executed as a five-step pipeline: The initial step (1) involves the extraction of raw data from two distinct sources (see Section 3.1). In the second step (2), we subject the data to a series of heuristic filters (see Section 3.2). We subsequently provide the refined data to DeepSeek-R1-0528, which generates instruction-answer pairs in Luxembourgish (3). Following this, we implement an LLM-as-a-judge approach to evaluate the quality of our synthetic data (4). In parallel, a subset of these pairs undergoes manual human evaluation. In the final step, we apply a post-filtering process to remove all samples that received poor scores (5), resulting in the final LuxIT dataset.

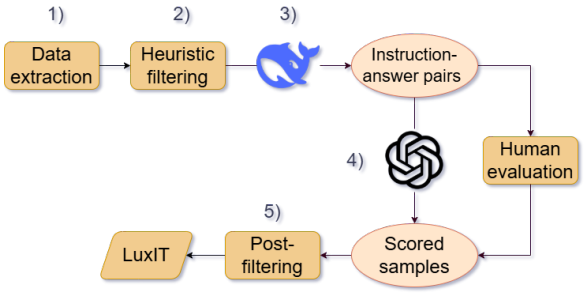


Figure 1: LuxIT Data Generation Pipeline.

3.1 Data

Our methodology utilizes two primary data sources: a complete dump of the Luxembourgish Wikipedia and a collection of all news articles and comments from RTL up to May 2024. We obtained the most recent dump of the Luxembourgish Wikipedia[4] on August 5th, 2025 and subsequently use Wikiextractor (Attardi, 2015) for extraction and cleaning. We convert the processed data from both sources into a JSON format for easier handling. Further details on the data structure are available in Appendix A.1. The initial collection comprises approximately 1.2 million RTL comments, 303,000 RTL news articles, and 80,000 Wikipedia articles, summing to approximately 1.6 million entries. We choose to exclude RTL user comments from our seed data due to their brevity and the potential for biased content, typographical mistakes, and grammatical inaccuracies.

3.2 Heuristic filtering

Before the generation phase, we implement a set of heuristic filters to remove low-quality samples. We retain only articles containing at least 750 characters. The threshold was set heuristically, to ensure each seed carries enough information. For Wikipedia articles, we exclude list articles, disambiguation pages, and pages marked as stubs. We filter out all non-Luxembourgish RTL news articles based on the language tags provided directly by RTL and apply additional filters, including boilerplate removal and semantic format filtering, which identifies keywords indicating a non-prose structure. This filtering process yields 16,558 Wikipedia articles and 100,340 RTL news articles.

3.3 Instruction-Answer pair generation

We employ DeepSeek-R1-0528[5] to synthetically create instruction-answer pairs in Luxembourgish from the filtered RTL news and Wikipedia articles. We select all 16,558 Wikipedia articles and randomly sample 66,776 RTL news articles. We then prompt the model to generate three instruction-answer pairs for each seed, yielding a total of 245,624 synthetic pairs[6].

To ensure the generation of high-quality data, we engineer a comprehensive prompt designed to process information from both data sources effectively. The prompt instructs the model to embed all necessary context within the instruction to facilitate a self-contained answer. For news articles, we

[4] https://dumps.wikimedia.org/lbwiki/latest/

[5] While Lothritz et al. (2026) used the original Deep-Seek-R1 in their experiments, we use the latest DeepSeek-R1-0528

[6] There were some parsing errors for the generated json strings, hence we exclude those samples.

direct the model to handle temporal context with care. We also guide the model to produce a diverse set of instruction types, including summarization, question answering, information extraction, and explanation, where the latter task consists of providing a justification or reasoning alongside the response. For summarization tasks specifically, the model is required to include the original source text within the instruction itself. The complete data generation prompt is available in Appendix A.2.

The final LuxIT dataset is structured with four columns: `instruction`, `response`, `messages` and `source_type` (RTL/Wiki). The `instruction` and `response` columns store the generated instructions and their corresponding answers. The `messages` column consolidates these two into ChatML format. We show the structure of our prompt template in Appendix A.4 and provide a representative sample of an instruction-answer pair from LuxIT in Table 4 in Appendix A.3.

3.4 LLM-as-a-judge scoring

Following generation, we apply an LLM-as-a-judge approach for quality control using GPT-5-mini (OpenAI, 2025). We select GPT-5-mini over GPT-4o-mini based on a comparison reported in Table 5 (Appendix A.5) which shows that despite a slight underperformance at lower Common European Framework of Reference for Languages (Council of Europe, 2001) (CEFR) levels, GPT-5-mini outperforms GPT-4o-mini by up to 24 points at higher levels (+9.7 pp macro average). A small-scale pilot comparing GPT-5 and GPT-5-mini judging found retention rates differing by only a few percent, justifying the cost-performance tradeoff.

For the judging task, we prompt GPT-5-mini to assign scores to each pair based on four criteria: linguistic quality, factual accuracy, instruction adherence (or instruction following), and helpfulness & relevance. Scoring is performed on the instruction–response pair in isolation. We use a three-point scoring system, with 1 representing poor quality and 3 indicating excellent quality. The linguistic quality score aims to check for the overall quality of the Luxembourgish language in the text, such as grammar, spelling and phrasing. Factual accuracy simply checks whether the answer factually answers the instruction without contradictions. Instruction adherence aims to verify whether the response correctly answers the instruction as intended and helpfulness relevance examines the general usefulness of the instruction. The full breakdown of our scoring system and the prompt used for this evaluation are provided in Appendices A.6 and A.7. We retain only samples that achieve a score of at least 2 across all four metrics, discarding all others to form the final dataset.

3.4.1 Human evaluation

To validate our automated quality control, we conduct a manual human evaluation on a random subset of 100 generated samples. This sample size is selected to provide a meaningful preliminary audit of the dataset's quality while remaining feasible within the standard resource constraints of low-resource language research, where recruiting qualified native-speaker annotators is both challenging and cost-intensive. The evaluation was performed independently by three native Luxembourgish speakers (two male, one female; ages 26, 32, and 35) who hold academic backgrounds in computer science and history.

Each annotator evaluated the samples using the identical 3-point scoring rubric applied by the automated LLM judge. Annotators were instructed to assess each pair using only the information contained in the instruction and response, without consulting the seed article. All scoring disagreements are resolved through a strict majority vote to determine the final aggregated label for each sample.

The automated judge retains a higher proportion of samples (92.6%) than the human evaluation (69.0%), indicating a leniency bias in the LLM judge. Inter-annotator agreement (Fleiss' Kappa) is fair for linguistic quality ($\kappa = 0.243$) but poor across the remaining metrics, including below-chance agreement for factual accuracy ($\kappa = -0.019$). This reflects the inherent difficulty of consistently evaluating synthetic Luxembourgish text and suggests that future annotation efforts would benefit from more granular rubrics and calibration sessions. Despite this low agreement, majority-vote aggregation over retained samples shows 94.2% rated "Excellent" for factual accuracy, pointing to generally high quality while underscoring the subjectivity of the individual judgments. We report the full breakdown in Table 7 in Appendix A.8.1.

To directly evaluate the alignment between our automated judge and human insight, Table 1 presents a per-item agreement analysis across the 100-sample subset. Exact agreement is low

Evaluation Metric	Exact Agreement	Cohen's κ	Human Mean (Std)	LLM Mean (Std)
Linguistic Quality	57.0%	0.121	2.24 (0.57)	2.19 (0.46)
Factual Accuracy	45.0%	-0.058	2.84 (0.39)	2.40 (0.65)
Instruction Adherence	83.0%	0.224	2.93 (0.26)	2.80 (0.45)
Helpfulness Relevance	51.0%	0.076	2.29 (0.88)	2.74 (0.44)

Table 1: Comparison of human and LLM evaluator (GPT-5-mini) agreement across four quality metrics on a random subset of 100 samples. Exact Agreement reports the percentage of matching scores, Cohen's κ measures inter-rater reliability, and Mean (Std) values are on a 1–3 scale.

and Cohen's Kappa reveals poor reliability overall, highlighted by a negative coefficient for Factual Accuracy ($\kappa = -0.058$). This below-chance agreement is consistent with a mismatch visible in the metric means. For Factual Accuracy, human reviewers heavily favored perfect marks, resulting in a highly skewed mean of 2.84. Conversely, the LLM judge was significantly more conservative, frequently downgrading those same pairs to acceptable, leading to a lower mean of 2.40. By contrast, a reverse trend is visible in Helpfulness Relevance, where the LLM displays a leniency bias, overestimating value compared to humans (2.74 vs. 2.29).

3.5 Post-filtering

After filtering out all samples with at least one score of 1, the final version of our dataset contains 227,507 instruction-answer pairs in Luxembourgish. We provide a detailed overview in Table 6 (Appendix A.8) and show the score distribution on LuxIT from our post-filtering step. The original dataset consists of 245,624 instruction-answer pairs. We reject 18,117 samples with low scores, observing a retention rate of 92.62%. Among all samples, 32,215 (13.12%) have perfect scores (score 3) across all score types. We observe that `instruction adherence` scores best among all metrics, with 223,065 (90.8%) obtaining a score of 3. Furthermore, among the rejected samples, `factual accuracy` has the highest rejection rate, with 11,745 entries having a score of 1. For the rejection rate, we observe 6,924, 11,745, 1,039 and 246 entries of score 1 (lowest) for linguistic quality, factual accuracy, instruction adherence and helpfulness relevance respectively.

3.6 Dataset composition

To characterize the task-type distribution of LuxIT and evaluate whether our open-ended generation prompt produces balanced coverage, we conduct a classification audit. While categorizing the entire 227,507-pair dataset is computationally prohibitive, we sample 100 random pairs and classify their instruction types using GPT-5.5 (OpenAI, 2026). The distribution of the sampled dataset is as follows: Summarization (32%), Information Extraction (30%), Question Answering (23%), and Explanation (15%). These results indicate that the generation prompt yields a relatively balanced mix of task types, though it exhibits a slight skew toward summarization and extraction tasks.

4 Experimental setup

In this section, we outline and break down our experiments in detail.

4.1 Fine-tuning on LuxIT

We fine-tune a variety of small models ($\leq$15B parameters) on LuxIT. We split the dataset into train, evaluation and test split with ratios 0.92, 0.05 and 0.03 respectively, training on 209,307 and evaluating on 11,375 samples. We select the following models: Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct (Grattafiori et al., 2024), Gemma-3-1B-IT, Gemma-3-12B-IT (Team et al., 2025), GLM-4-9B-0414 (GLM et al., 2024), Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct (Qwen et al., 2025), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Ministral-3-3B-Instruct-2512 (Liu et al., 2026), Phi-4 (Abdin et al., 2024), Apertus-8B-Instruct-2509 (Hernández-Cano et al., 2026), EuroLLM-1.7B-Instruct (Martins et al., 2025), Olmo-3-7B-Instruct (Olmo et al., 2025). The models are fine-tuned with LoRA (Hu et al., 2022) using the Unsloth[7] framework. Unless otherwise specified, we apply a standard configuration across models: a LoRA rank (r) of 16, a LoRA alpha (α) of 32, a learning rate of 1×10^{-4} with a cosine learning rate scheduler, and a warmup ratio of 0.03. Models are trained for 2 epochs using

[7] https://docs.unsloth.ai/

the AdamW 8-bit optimizer. Due to computational constraints, hardware setups differ slightly between runs. A comprehensive table detailing the hyperparameters, exceptions, and GPU hardware used for each specific model can be found in Appendix B.1.

4.2 Evaluation on language exams

We investigate the impact of LuxIT on the Luxembourgish language proficiency of small LLMs, following the same approach proposed by Lothritz et al. (2026). After fine-tuning, we evaluate the LLMs on Luxembourgish language proficiency exams, instructing the models to solve exams consisting of multiple-choice questions, testing for vocabulary, grammar, reading comprehension, and conversational comprehension (done through transcripts). The language exams stem from the Luxembourgish language institute *Institut National des Langues Luxembourg (INLL)*[8] and are divided into 6 skill levels defined in the CEFR, ranging from A1 (basic level) to C2 (native level), each exam level consisting of fill-in-the-blank and multiple choice questions (Lothritz et al., 2026). The full dataset consists of 629 such questions, with slightly more than 100 questions per difficulty level and a nearly equal distribution of testing categories (vocabulary, grammar, reading comprehension, conversational comprehension). We measure the models' performance using accuracy as a metric, akin to Lothritz et al. (2026). We compare instruct models to our fine-tuned models.

The INLL exams are proprietary and not publicly accessible. This substantially reduces the risk of data contamination within our Wikipedia and RTL news seed corpora, ensuring that the models' performance reflects genuine language comprehension rather than test-set memorization.

4.3 Evaluation on Luxembourgish downstream tasks

We evaluate our 14 fine-tuned models on 5 well-known NLP downstream tasks in Luxembourgish, namely Intent Classification (IC) (Lothritz et al., 2021), Winograd Natural Language Inference (WNLI) (Lothritz et al., 2022), Recognizing Textual Entailment (RTE), Sentiment Analysis and Stanford Sentiment Treebank (SST-2), taken from Lothritz et al. (2023). According to the authors, the data in these task-benchmarks was not taken from Wikipedia or RTL articles. The dataset consists of 3,823 samples. We report the class distribution in Table 9 in Appendix B.2.

[8] https://www.inll.lu/en/

5 Results

5.1 RQ1 i): Does fine-tuning on LuxIT improve overall accuracy across CEFR proficiency levels?

Table 2 shows the comparison between our fine-tuned models and their respective base models evaluated on Luxembourgish language exams, reporting total accuracy on all language exam levels. Fine-tuning on LuxIT leads to a macro average accuracy increase in 12 out of 14 models, with macro average gains ranging from +2.5 percentage points (pp) (Gemma-3-12B-IT) to +21.0 pp (Ministral-3-3B-Instruct). On average, the models see a macro average accuracy increase of 5.37 pp. The effectiveness of instruction tuning on LuxIT is most evident for Ministral-3-3B-Instruct, with its macro average increasing from 25.8% to 46.8% (+21.0 pp). Similarly, for Llama-3.2-1B-Instruct, Phi-4 and Qwen2.5-7B-Instruct, the macro average increases by 14.4 pp, 14.2 pp and 13.6 pp respectively. For 2 of the 14 models tested, Gemma-3-1B and Qwen2.5-0.5B, we observe a degradation. We observe a decrease in macro average accuracy of -22.0 pp and -13.6 pp, respectively. With regard to the individual language exam levels, we observe the most consistent increase for B1/B2 level across all models except our 2 overall decreasing models, Gemma-3-1B-IT and Qwen2.5-0.5B-Instruct. C2 level shows the most inconsistent gains, with 6 out of 14 models declining.

Out of all models, Gemma-3-12B-IT has the strongest baseline before fine-tuning, with a baseline macro average of 56.9%. We identify GLM-4-9B-0414 as the top performing model in terms of absolute performance, with a final macro average of 60.6%, followed by Gemma-3-12B-Instruct with an average of 59.4%. Appendix C.1 contains a more detailed visual breakdown.

5.2 RQ1 ii): How does the impact vary across individual linguistic categories?

Table 3 shows macro-averaged accuracy changes across the four skill categories introduced in section 4.2. The largest gains are observed in vocabulary and conversational comprehension, where multiple models improve substantially — notably Llama-3.2-1B-Instruct (+24.4 pp vocabulary) and Phi-4 (+25.8 pp conversational comprehension). Read-

Model	A1	A2	B1	B2	C1	C2	Macro Avg
	LuxIT (Δbase)	LuxIT (Δbase)	LuxIT (Δbase)	LuxIT (Δbase)	LuxIT (Δbase)	LuxIT (Δbase)	LuxIT (Δbase)
∼0.5–1.5B parameter models							
qwen2.5-0.5b	20.2 (−22.1)	21.2 (−9.6)	12.6 (−16.5)	23.7 (−8.8)	15.4 (−3.9)	6.9 (−20.8)	16.7 (−13.6)
qwen2.5-1.5b	41.3 (+1.0)	40.4 (+1.0)	34.0 (+10.7)	38.6 (+15.8)	29.8 (+1.0)	27.7 (+6.9)	35.3 (+6.0)
llama-3.2-1b	36.5 (+17.3)	32.7 (+21.2)	24.3 (+10.7)	26.3 (+14.0)	21.2 (+11.5)	22.8 (+11.9)	27.3 (+14.4)
gemma-3-1b	23.1 (−22.1)	21.2 (−18.3)	13.6 (−24.3)	10.5 (−25.4)	7.7 (−18.3)	5.9 (−23.8)	13.7 (−22.0)
eurollm-1.7b	26.9 (±0.0)	26.9 (+3.8)	22.3 (+2.9)	27.2 (+7.9)	14.4 (±0.0)	21.8 (+8.9)	23.3 (+3.9)
ministral-3-3b	61.5 (+25.0)	48.1 (+16.3)	45.6 (+22.3)	48.2 (+29.8)	38.5 (+13.5)	38.6 (+18.8)	46.8 (+21.0)
∼7–9B parameter models							
mistral-7b	51.9 (+3.8)	45.2 (+5.8)	41.7 (+13.6)	44.7 (+10.5)	29.8 (+12.5)	26.7 (+4.9)	40.0 (+8.5)
olmo-3-7b	46.2 (+14.4)	48.1 (+23.1)	38.8 (+10.7)	33.3 (+2.6)	35.6 (+12.5)	22.8 (−4.0)	37.5 (+9.9)
qwen2.5-7b	73.1 (+21.2)	59.6 (+14.4)	56.3 (+16.5)	61.4 (+15.8)	48.1 (+14.4)	36.6 (−1.0)	55.9 (+13.6)
llama-3.1-8b	49.0 (±0.0)	46.2 (+1.9)	41.7 (+10.7)	40.4 (+1.8)	39.4 (+7.7)	25.7 (−4.0)	40.4 (+3.0)
apertus-8b	56.7 (+3.8)	51.0 (−1.0)	48.5 (+13.6)	46.5 (+10.5)	36.5 (+2.9)	32.7 (+3.0)	45.3 (+5.5)
glm-4-9b	74.0 (+11.5)	65.4 (+2.9)	59.2 (+11.6)	59.6 (+7.0)	53.8 (+7.7)	51.5 (+8.9)	60.6 (+8.3)
∼10–14B parameter models							
gemma-3-12b	68.3 (±0.0)	64.4 (±0.0)	64.1 (+8.7)	60.5 (+3.5)	58.7 (+5.8)	40.6 (−3.0)	59.4 (+2.5)
phi-4	67.3 (+7.7)	65.4 (+16.3)	58.3 (+20.4)	52.6 (+7.0)	64.4 (+24.0)	46.5 (+9.9)	59.1 (+14.2)

Table 2: Accuracy (%) of LuxIT fine-tuned models on Luxembourgish language exams, with change relative to the base model (pre-fine-tuning instruct model) in brackets. Green indicates improvement over the base, red indicates degradation. Δbase = macro-averaged accuracy of the LuxIT fine-tuned model minus that of the base model (higher is better).

Model	Vocabulary		Grammar		Reading		Conversation	
	LuxIT	ΔBase	LuxIT	ΔBase	LuxIT	ΔBase	LuxIT	ΔBase
∼0.5–1.5B parameter models								
qwen2.5-0.5b	26.2	(−4.1)	26.2	(−5.1)	5.1	(−25.6)	9.1	(−19.9)
qwen2.5-1.5b	36.8	(+9.6)	33.3	(+2.4)	33.3	(+2.6)	38.0	(+9.0)
llama-3.2-1b	37.5	(+24.4)	32.6	(+8.8)	16.7	(+7.7)	21.9	(+16.8)
gemma-3-1b	14.8	(−24.6)	14.0	(−24.1)	10.9	(−21.8)	14.8	(−17.4)
eurollm-1.7b	29.0	(+5.8)	29.9	(+2.0)	14.1	(+3.2)	19.3	(+4.5)
ministral-3-3b	54.7	(+15.9)	44.9	(+11.8)	43.6	(+30.8)	44.5	(+25.9)
∼7–9B parameter models								
mistral-7b	47.6	(+20.1)	33.5	(+1.9)	41.0	(+8.3)	38.6	(+4.6)
olmo-3-7b	41.6	(+14.0)	29.6	(+1.0)	39.1	(+13.5)	40.0	(+10.9)
qwen2.5-7b	65.8	(+14.1)	40.0	(+8.2)	59.0	(+13.5)	59.3	(+18.2)
llama-3.1-8b	48.5	(+7.9)	33.4	(+1.8)	39.7	(+4.5)	40.7	(−1.8)
apertus-8b	53.4	(−2.2)	33.9	(+8.5)	44.2	(+4.5)	50.9	(+11.0)
glm-4-9b	74.6	(+3.1)	42.8	(+7.1)	56.4	(+7.7)	69.6	(+15.5)
∼10–14B parameter models								
gemma-3-12b	70.5	(−1.1)	41.1	(+1.9)	59.6	(+4.5)	67.7	(+5.1)
phi-4	68.6	(+18.3)	43.4	(+6.0)	54.5	(+7.7)	71.6	(+25.8)

Table 3: Macro-averaged accuracy (%) per skill category for each LuxIT fine-tuned model, together with the change over the respective base model. Green = improvement; red = degradation. ΔBase = LuxIT accuracy minus base accuracy (higher is better).

ing comprehension shows more variable results, with large gains for some models (Ministral-3-3B-Instruct: +30.8 pp) but modest improvements elsewhere. Grammar scores show the most modest changes, with gains rarely exceeding 10 pp, suggesting that the synthetic data is less effective at improving structural linguistic knowledge. The two degrading models (Qwen2.5-0.5B-Instruct and Gemma-3-1B-IT) show uniform drops across all four categories, consistent with their overall regression on language exams in Table 2.

5.3 RQ2: Does fine-tuning on LuxIT improve performance on downstream Luxembourgish NLP tasks?

Figure 2 shows the average change in performance on the Luxembourgish NLP tasks (Section 4.3), comparing baseline and LuxIT fine-tuned models (see Appendix C.2 for a breakdown of the results). Fine-tuning on LuxIT leads to improvements in Luxembourgish NLP task performance for 9 out of 14 models, as measured by macro-averaged F1 (Figure 2) across intent classification, RTE, sentiment analysis, SST-2, and WNLI. The most substantial gains are observed for Phi-4 (ΔF1-macro = +0.325), Olmo-3-7b (+0.211), and qwen2.5-7b (+0.172), while the largest regressions occur for Apertus-8b (-0.206) and Gemma-3-12b (-0.128). Models that improve on language exams (RQ1) do not systematically improve on NLP tasks, and vice versa with notable divergences for Ministral-3-3B, which shows the largest language exam gain (+21.0pp) alongside near-zero NLP task change (-0.003), and Gemma-3-1b, which shows a modest NLP task gain (+0.064) despite a substantial language exam regression (-22.0pp).

It is to note that there are substantial outliers that influence the average performance change (see Appendix C.2, Figure 4). Most notably, Apertus-8b performs poorly on the RTE task after fine-tuning on LuxIT, with a change in performance of -0.677 and Gemma3-12b's performance on IC changes by

-0.522. On the other hand, we observe a considerable gain of +0.802 for Phi-4 in the SST task. While those outliers explain the extreme average changes we observe for Phi-4, Gemma3-12b, and Apertus-8b, it is unclear why those outliers occur.

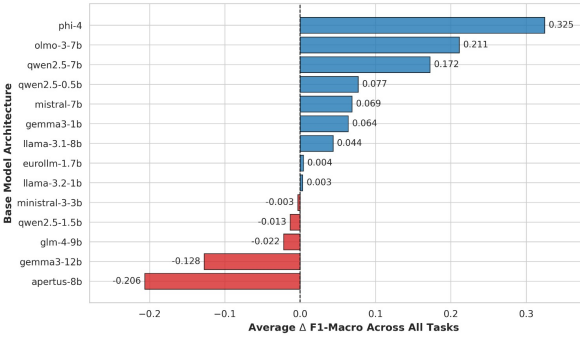


Figure 2: Average Impact of LuxIT Fine-Tuning per Model across all tasks.

6 Discussion

LuxIT represents an important step toward addressing the scarcity of instruction-tuning resources for Luxembourgish. Our pipeline effectively synthesizes a large dataset from monolingual seed data. Furthermore, our dataset composition audit demonstrates that the open-ended generation prompt yields a balanced coverage of diverse instruction types (summarization, extraction, Q&A, and explanation). The human evaluation underscores the subjectivity of manually evaluating synthetic Luxembourgish text. We observe low human inter-annotator agreement (Table 7) alongside metric-specific biases between the automated judge and native speakers (Table 1). While the LLM judge shows a leniency bias regarding Helpfulness Relevance, human evaluators are more forgiving regarding Factual Accuracy. This divergence highlights the need for more objective evaluation frameworks for future iterations.

Regarding language proficiency, fine-tuning generally enhanced formal Luxembourgish comprehension. However, the divergent outcomes among the smallest models suggest that fine-tuning receptivity depends on more than model size alone. Models with limited prior task alignment saw the most benefit, whereas regressions in other sub-1.5B models with stronger baselines likely point to catastrophic forgetting of pre-trained representations.

A deeper analysis of the language exam levels reveals that gains from LuxIT are not uniformly distributed across proficiency levels. Improvements were most consistent at intermediate proficiency levels (B1/B2) and in vocabulary and conversational comprehension. Contrarily, structural grammar and advanced native-like nuances (C2) proved difficult to improve. This suggests that the synthetic data, constrained by the generator model's capabilities and the formal nature of the seed corpora, effectively expands the models' lexicon but skews toward intermediate proficiency.

Furthermore, the limited correlation between language exam improvements and downstream NLP task performance highlights the specific nature of our dataset. The language exams feature standardized, well-formed Luxembourgish, while the NLP tasks contain noisy, non-standardized text. Because LuxIT is derived from formal Wikipedia and news articles, the fine-tuned models appear to align better with the language style of the formal exams. This divergence suggests that format alignment and linguistic improvement are separable effects that instruction tuning on LuxIT does not uniformly induce across all models.

Despite these nuances, in terms of absolute capability, fine-tuning medium-sized models on LuxIT successfully yielded highly proficient Luxembourgish models.

7 Conclusion

This work addresses the scarcity of instruction tuning resources for low-resource languages by introducing LuxIT, a high-quality dataset synthesized from monolingual Luxembourgish seed texts entirely within the target language. By fine-tuning fourteen small-scale LLMs, we demonstrate that LuxIT broadly improves formal language comprehension, particularly in expanding vocabulary and enhancing intermediate-level proficiency, even if performance on noisy, structured downstream NLP tasks remains less consistent. Ultimately, our findings validate that leveraging monolingual synthetic data is an effective and feasible strategy for enhancing LLM capabilities in low-resource linguistic settings. Future work will explore open-ended conversational evaluations and multi-turn data integration.

Limitations

We acknowledge several limitations. First, LuxIT is derived from Wikipedia and RTL news articles and therefore skewed toward formal, standardized, and objective language registers. Consequently, LuxIT may lack the conversational flow, regional

dialectal variations, or informal creative language styles of everyday Luxembourgish. Future data synthesis efforts should incorporate conversational transcripts or informal text.

Second, our human evaluation, while crucial for validation, was conducted on a relatively small subset of 100 samples. A larger-scale human evaluation by multiple native speakers would lend greater statistical power to our quality claims and provide deeper insights into the dataset's strengths and weaknesses.

Third, we acknowledge the observed leniency bias of our single LLM judge (GPT-5-mini) and human evaluators. As shown by the low Exact Agreement and Cohen's Kappa scores (Table 1), biases are metric-dependent. The LLM overestimates Helpfulness Relevance compared to humans, suggesting the final dataset may contain instructions that are marginally less useful than their automated scores imply. Furthermore, the human audit reveals a negative inter-annotator agreement among the native speakers themselves (Fleiss' κ = -0.019, Table 7). Combined with their skewed leniency compared to the LLM (human mean of 2.84 vs. the LLM's 2.40), this indicates our evaluation guidelines allowed for highly subjective interpretations. This emphasizes the difficulty of manually evaluating synthetic text in Luxembourgish. Future iterations should implement more granular rubrics and annotator calibration to improve human consistency, alongside stricter threshold prompting or multi-model LLM ensembles to enhance automated filtering. Similarly, our factual accuracy rubric should not be interpreted as a faithfulness metric. With the exception of the cases where the generation model reproduced the seed text within the instruction, neither the automated judge nor the human annotators were shown the source article. A response that is fluent and plausible but untrue with respect to the seed article would therefore not reliably be filtered out.

Fourth, while our approach is strictly monolingual in its seed data, the generation model, DeepSeek-R1-0528, is a multilingual model. Its internal knowledge is shaped by the vast multilingual corpus it was pre-trained on, which could subtly influence the style and structure of the generated Luxembourgish. More fundamentally, our pipeline requires a generator model already proficient in the target language.

Fifth, our downstream evaluation is restricted to multiple-choice language proficiency exams and classification tasks. While this provides a rigorous and objective measure of grammatical, lexical, and comprehension skills, it does not evaluate the models' open-ended text generation, multi-turn conversational abilities, or specific instruction-following nuances. Consequently, the true conversational proficiency of the fine-tuned models remains to be evaluated.

Finally, Apertus-8B-Instruct is constrained to a maximum sequence length of 512 tokens due to hardware limitations, versus 2048 tokens for all other models. As some LuxIT pairs may exceed this length, its results may underestimate the model's true capacity under standard training conditions.

Ethical considerations and licensing

The Luxembourgish Wikipedia seed data is available under the CC BY-SA 4.0 license, and our use and distribution of RTL news articles is governed by a formal data-sharing agreement with RTL. As DeepSeek-R1-0528 places no restrictions on its outputs, we release the LuxIT Wikipedia subset under CC BY-SA 4.0 and the RTL subset under CC BY-NC 4.0. The INLL proficiency exams (Section 4.2) are proprietary and not publicly distributed; we obtained explicit permission from INLL to use them solely for the automated evaluation of our models. The five downstream NLP benchmarks are used under their original academic terms.

Synthesizing data with LLMs risks propagating biases present in the seed texts or the generator model. We therefore excluded user-generated RTL comments, which are more prone to biased and toxic content, and applied heuristic filtering favoring formal, objective language.

Acknowledgments

This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470 , and by the BESSER project through the Luxembourg National Research Fund (FNR) PEARL programme under the grant agreement 16544475. Our gratitude goes out to RTL for supplying the Luxembourgish news articles and to INLL for providing the Luxembourgish language exams. We thank Aline Fries and Aaron Conrardy for helping us with the human evaluation part and Bernhard Voggenberger for advising us on the fine-tuning experiments.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. 2024. Phi-4 technical report. Preprint, arXiv:2412.08905.

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, et al. 2023. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. Accessed on July 31, 2026.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge, UK.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, et al. 2026. Apertus: Democratizing open and compliant LLMs for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. 2023. Mistral 7b. Preprint, arXiv:2310.06825.

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.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, et al. 2023a. Mimic-it: Multi-modal in-context instruction tuning. Preprint, arXiv:2306.05425.

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

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, et al. 2026. Ministral 3. Preprint, arXiv:2601.08584.

Cedric Lothritz, Kevin Allix, Bertrand Lebichot, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. 2021. Comparing multilingual and multiple monolingual models for intent classification and slot filling. In *Natural Language Processing and Information Systems*, pages 367–375, Cham. Springer International Publishing.

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, Saad Ezzini, Christoph Purschke, Tegawendé Bissyandé, Jacques Klein, Isabella Olariu, et al. 2023. Comparing pre-training schemes for Luxembourgish BERT models. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 17–27, Ingolstadt, Germany. Association for Computational Lingustics.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, et al. 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.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62. Proceedings of the Second EuroHPC user day.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, et al. 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.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL*

2024, pages 12585–12611, Bangkok, Thailand. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, et al. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, et al. 2025. Olmo 3. Preprint, arXiv:2512.13961.

OpenAI. 2025. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/. Accessed on August 18th, 2025.

OpenAI. 2026. Gpt-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf. Accessed on March 6th, 2026.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. Preprint, arXiv:2304.03277.

Fred Philippy, Laura Bernardy, Siwen Guo, Jacques Klein, and Tegawendé F. Bissyandé. 2025a. Luxinstruct: A cross-lingual instruction tuning dataset for luxembourgish. Preprint, arXiv:2510.07074.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawende Bissyande. 2025b. 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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2025. Qwen2.5 technical report. Preprint, arXiv:2412.15115.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 61836–61856. Curran Associates, Inc.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, et al. 2023. Alpaca: A strong, replicable instruction-following model. Accessed on September 16, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al. 2025. Gemma 3 technical report. Preprint, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, et al. 2023a. Llama: Open and efficient foundation language models. Preprint, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. Preprint, arXiv:2307.09288.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, et al. 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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, et al. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, et al. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems*, volume 36, pages 26650–26685. Curran Associates, Inc.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, et al. 2025. Instruction tuning for large language models: A survey. Preprint, arXiv:2308.10792.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, et al. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A LuxIT

A.1 Data structure

Both data sources are contained in a JSON file. The following shows an example entry of an

RTL news article[9] and a Wikipedia article. We only extract `public_date`, `title`, `header` and `text` from the RTL news article and extract `title` and `text` from the Wikipedia article.

RTL news article


```
{
 'category_name': 'International',
 'category_id': 5,
 'article_id': 2190645,
 'type': 'news',
 'public_date': '2024-04-28 14:04:54',
 'title': '<article title>',
 'header': '<article header>',
 'text': '<article text>',
 'tags': ['international'],
 'text_id': 301779,
 'lang_id': 'lb'
}
```


Wikipedia article


```
{
  "id": '25',
  "revid": '580',
  "url": "https://lb.wikipedia.org/
    wiki?curid=25",
  "title": "Matthew Perry",
  "text": "De Matthew Langford Perry,
    gebuer den 19. August 1969 zu
    Williamstown am Massachusetts, a
    gestuerwen den 28. Oktober 2023
    zu Los Angeles, war en
    US-amerikanesch-kanadesche
    Schauspiller, dee virun allem
    duerch seng Roll als Chandler
    Muriel Bing an der
    Televisiounsserie Friends bekannt
    ginn ass..."
}
```


A.2 Data generation model prompt

The data generation is performed by DeepSeek-R1-0528. We instruct the model to return the data in JSON format, to ensure compatible formatting and to make it easily accessible. The prompt is formulated as follows:

Data Generation Model Prompt

```
You are an expert in the
    Luxembourgish language tasked
    with creating high-quality
    synthetic training data for
    language models.

OBJECTIVE:
Generate 3 instruction-response
    pair(s) in authentic
    Luxembourgish based on the
    provided text. These pairs will
    be used for instruction
    fine-tuning of language models.

REQUIREMENTS:
1. LANGUAGE: All content MUST be in
    fluent, natural Luxembourgish
   - Use proper Luxembourgish
    grammar, spelling, and idioms
   - Avoid unnecessary German or
    French loan words
   - Ensure the language sounds
    natural to native speakers

2. QUALITY STANDARDS:
   - Instructions should be clear,
    specific, and answerable based on
    the provided text
   - Responses should be
    comprehensive, accurate, and
    well-structured
   - Include ALL necessary context in
    the instruction for a complete
    answer
   - If insufficient information
    exists, indicate that more
    details are needed

3. SUMMARIZATION INSTRUCTIONS:
   - When creating summarization
    tasks, ALWAYS include the
    original seed-text unchanged in
    the instruction for reference
   - The instruction should present
    the source text and ask for a
    summary
   - This ensures the training data
    contains both the source material
    and the summary

4. TEMPORAL CONTEXT:
   - When a date is provided,
    incorporate it appropriately
   - Add temporal context to maintain
    relevance when applicable
   - Consider whether dates belong in
    the instruction, response, or both

5. DIVERSITY:
   - Create varied types of
    instructions (e.g.,
    summarization, Q&A, information
    extraction, explanation)
   - Vary complexity levels
    appropriately
   - Ensure each pair is unique and
    adds value

6. OUTPUT FORMAT:
   Return ONLY a valid JSON array
    with the following structure:
   [
     {{
       "instruction": "Clear
    instruction in Luxembourgish",
```

[9] We do not show title, header and text for the RTL news article as we do not redistribute raw articles in this paper

```
      "response": "Detailed response
    in Luxembourgish"
      }}
   ]

SOURCE TEXT:
{source_context}

Generate 3 high-quality
    instruction-response pair(s)
    based on the above text.
```

A.3 LuxIT Sample

Here we show a sample of LuxIT taken from the Wikipedia subset.

Instruction	Response
Wien dierft bei de Gemengewalen wielen (aktiivt Walrecht), a wat sinn d'Konditioune fir Net-Lëtzebuerger?	Fir ze wielen, muss een 18 Joer al sinn, bierger-lech a politesch Rechter hunn an zu Lëtzebuerg wunnen. Net-Lëtzebuerger mussen hire Wunnsëtz zu Lëtzebuerg hunn, op der Gemeng an d'Wielerlëscht ageschriwwen sinn, a bei Bierger vu Länner baussent EU/Schengen e gültegt Openthaltsdokument hunn.

Table 4: Example instruction-response pair from the LuxIT dataset, taken from the Wikipedia subset.

A.4 Prompt Template

Here we show the prompt template in more detail that unifies columns from LuxIT (Table 4). The `Instruction` and `Response` columns are brought together in the ChatML format:

Prompt Template

```
messages =
[
    {
        "role": "user",
        "content": <instruction>
    },
    {
        "role": "assistant",
        "content": <response>
    }
]
```

A.5 LLM-as-a-judge

Here we break down the reasoning for using GPT-5-mini as a judge (Section 3.4). Table 5 shows the evaluation results from GPT-5 and GPT-5-mini compared to GPT-4o and GPT-4o-mini, where the accuracy scores for the latter 2 were taken from Lothritz et al. (2026) and we simply added the macro average.

Model	A1	A2	B1	B2	C1	C2	Avg
GPT-4o	92.3	85.6	**78.6**	84.2	74.0	60.4	79.2
GPT-4o-mini	90.4	90.4	59.2	59.6	54.8	43.6	66.3
GPT-5	**94.2**	**92.3**	**78.6**	**87.7**	**87.5**	**64.4**	**84.1**
GPT-5-mini	83.7	82.7	76.7	78.9	78.8	55.4	76.0
Δ *(5-mini vs. 4o-mini)*	−6.7	−7.7	+17.5	+19.3	+24.0	+11.8	+9.7

Table 5: Accuracy (%) on Luxembourgish CEFR-level proficiency exams (A1–C2). **Bold** denotes the best score per column. Shading marks our chosen judge model. The Δ row shows the gain/degradation of GPT-5-mini over GPT-4o-mini.

A.6 Custom scoring metric

We fully break down our custom scoring metric introduced in section 3.4:

Linguistic Quality score

Score 1: Contains significant grammatical errors, spelling mistakes, or unnatural phrasing in Luxembourgish. Text that is actually German or French instead of proper Luxembourgish should receive this score.

Score 2: Mostly correct Luxembourgish, but has minor errors or sounds slightly robotic/unnatural. May mix in too many loan words unnecessarily.

Score 3: Fluent, idiomatic, and grammatically perfect Luxembourgish. Natural-sounding text that a native speaker would produce.

Factual Accuracy score

Score 1: Contains factual errors that contradict the source text or general knowledge.

Score 2: Mostly accurate but might have minor inaccuracies or omissions.

Score 3: Completely accurate according to the source text and factual knowledge.

Instruction Adherence score

Score 1: Fails to follow the core instruction (e.g., provides a summary when asked for a list).

Score 2: Follows the main instruction but misses a constraint (e.g., writes 4 bullet points when asked for 3, wrong format, or incorrect tone).

Score 3: Perfectly follows all parts of the instruction, including constraints like length, format, and tone.

Helpfulness Relevance score

Score 1: The instruction is nonsensical, irrelevant to any reasonable context, or the response is unhelpful/off-topic.

Score 2: The instruction is plausible but not very insightful or creative. The response addresses the instruction but in a basic way.

Score 3: A genuinely useful, interesting, or creative instruction that elicits a helpful, comprehensive response.

A.7 Post-filtering model prompt

We post filter the synthetic data with GPT-5-mini. Again, we want the scores to be in JSON format. The prompt is as follows:

Post-Filtering Model Prompt

```
You are an expert evaluator of
    Luxembourgish text quality. Your
    task is to evaluate the following
    instruction-response pair written
    in Luxembourgish based on four
    specific criteria.

IMPORTANT: The texts below are in
    Luxembourgish. You must evaluate
    them as Luxembourgish texts, NOT
    as German, French, or any other
    language. Luxembourgish has its
    own distinct grammar, vocabulary,
    and spelling conventions.

INSTRUCTION (in Luxembourgish):
{instruction}

RESPONSE (in Luxembourgish):
{response}

EVALUATION CRITERIA:

1. linguistic_quality (Linguistic
    Quality):
   - Score 1 (Poor): Contains
    significant grammatical errors,
    spelling mistakes, or unnatural
    phrasing in Luxembourgish. Text
    that is actually German or French
    instead of proper Luxembourgish
    should receive this score.
   - Score 2 (Acceptable): Mostly
    correct Luxembourgish, but has
    minor errors or sounds slightly
    robotic/unnatural. May mix in too
    many loan words unnecessarily.
   - Score 3 (Excellent): Fluent,
    idiomatic, and grammatically
    perfect Luxembourgish.
    Natural-sounding text that a
    native speaker would produce.

2. factual_accuracy (Factual
    Accuracy):
   - Score 1 (Incorrect): Contains
    factual errors that contradict
    the source text or general
    knowledge.
   - Score 2 (Mostly Correct): Mostly
    accurate but might have minor
    inaccuracies or omissions.
   - Score 3 (Perfect): Completely
    accurate according to the source
    text and factual knowledge.

3. instruction_adherence (Instruction
    Following):
   - Score 1 (Not Followed): Fails to
    follow the core instruction
    (e.g., provides a summary when
    asked for a list).
   - Score 2 (Partially Followed):
    Follows the main instruction but
    misses a constraint (e.g., writes
    4 bullet points when asked for 3,
    wrong format, or incorrect tone).
   - Score 3 (Fully Followed):
    Perfectly follows all parts of
    the instruction, including
    constraints like length, format,
    and tone.

4. helpfulness_relevance (Helpfulness
    and Relevance):
   - Score 1 (Not Helpful): The
    instruction is nonsensical,
    irrelevant to any reasonable
    context, or the response is
    unhelpful/off-topic.
   - Score 2 (Somewhat Helpful): The
    instruction is plausible but not
    very insightful or creative. The
    response addresses the
    instruction but in a basic way.
   - Score 3 (Very Helpful): A
    genuinely useful, interesting, or
    creative instruction that elicits
    a helpful, comprehensive response.

CRITICAL INSTRUCTIONS:
- Respond ONLY with a JSON object
    containing the four scores.
- Each score must be an integer: 1,
    2, or 3.
- Do NOT include any explanations,
    comments, or additional text
    outside the JSON.
- Evaluate the text AS LUXEMBOURGISH,
    not as any other language.

JSON FORMAT:
{{
    "linguistic_quality": <score 1-3>,
    "factual_accuracy": <score 1-3>,
```

```
    "instruction_adherence": <score
    1-3>,
    "helpfulness_relevance": <score
    1-3>
}}
```

A.8 Dataset Quality

In Table 6 we provide detailed statistics on the LuxIT score distribution introduced in Section 3.5.

A.8.1 Human Evaluation

We provide details for our human evaluation results in Table 7.

B Experimental Setup

Here we provide more details accompanying the Experimental Setup Section (Section 4).

B.1 Fine-Tuning on LuxIT

We provide more details for the hyperparameter settings in Table 8. (Introduced in Section 4). Several hyperparameters and configurations are held constant across all models unless otherwise specified.

All models are fine-tuned for 2 epochs using the 8-bit AdamW optimizer and a cosine learning rate scheduler. We consistently apply a warmup ratio of 0.03, a weight decay of 0.01, and keep the default value for max_grad_norm at 1.0. For the Low-Rank Adaptation (LoRA) setup, the rank (r) is set to 16, the scaling factor (α) to 32, and the dropout rate to 0.05. Gradient checkpointing is enabled across all training runs to optimize memory consumption. During training, logging occurs every 200 steps, while evaluation and model checkpoint saving are performed every 500 steps. For the dataset, 5% is strictly held out for validation and 3% for testing across all runs.

Notably, while a maximum sequence length of 2048 tokens is utilized for the vast majority of the models, we constrain the Apertus-8B-Instruct model to a sequence length of 512 tokens due to hardware memory limitation requirements. Furthermore, all models trained via the Unsloth framework utilize a fixed random seed of 3407 to ensure reproducibility, whereas the Apertus-8B-Instruct model, which we train using standard Hugging Face pipelines, is initialized with a random seed of 42.

B.2 Evaluation on Luxembourgish downstream tasks

We break down the dataset class distribution in Table 9 introduced in Section 4.3.

Task	Samples
Intent Classification (IC)	159
Recognizing Textual Entailment (RTE)	801
Sentiment Analysis (SA)	367
Stanford Sentiment Treebank (SST-2)	2,360
Winograd NLI (WNLI)	136

Table 9: Sample distribution across Luxembourgish downstream tasks.

C Results

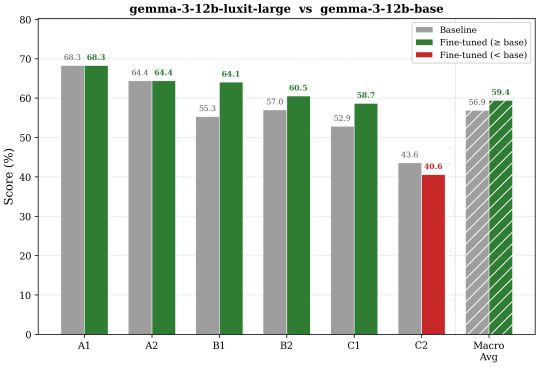
Here we provide more details accompanying the Results section (Section 5).

C.1 RQ1 i)

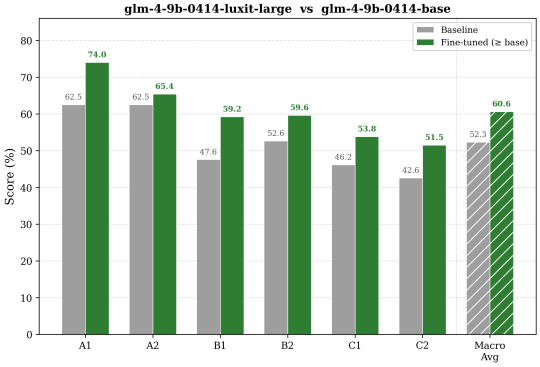
We provide detailed visual comparison between our fine-tuned models and their respective baseline in Figure 3 (Section 5.1).

C.2 RQ2

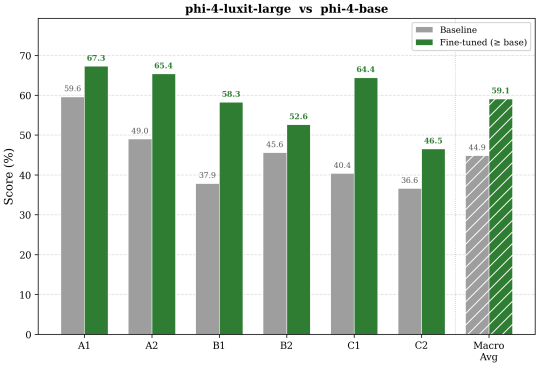
We report the results of benchmarking our fine-tuned models against their respective baseline models (Section 5.3) on the Luxembourgish downstream tasks in Table 10 (introduced in Section 4.3) and show the per-task changes and the average change per model in Figure 4 and Figure 2 respectively.



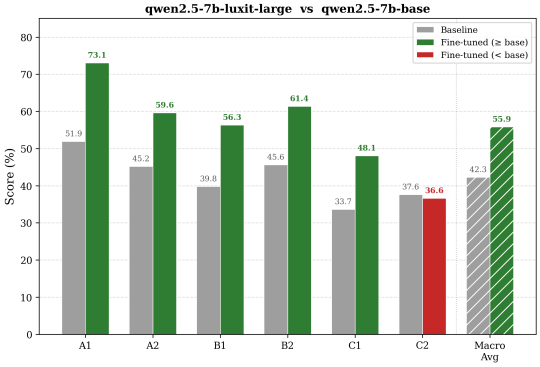
(a) gemma-3-12b



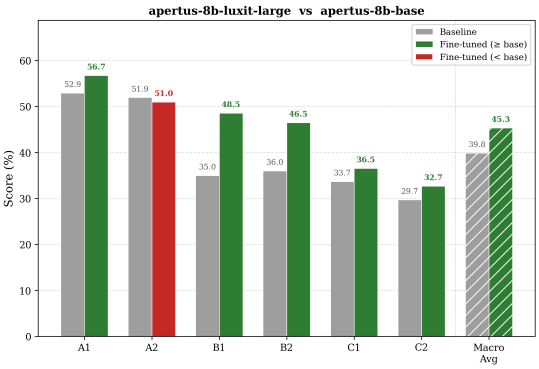
(b) glm-4-9b-0414



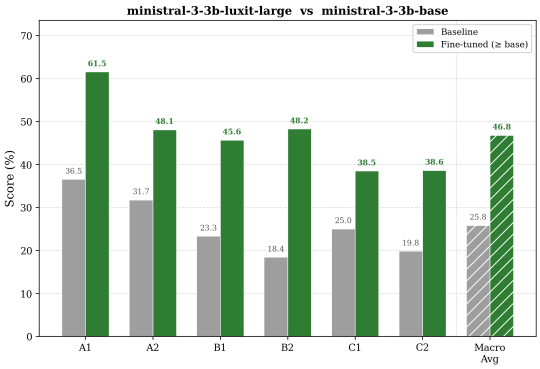
(c) phi-4



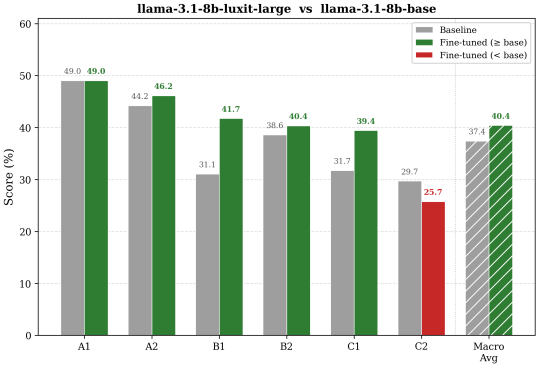
(d) qwen2.5-7b



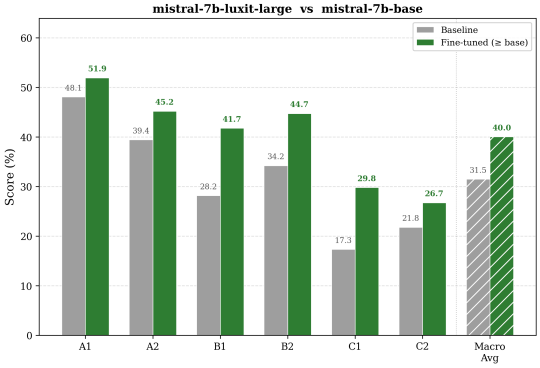
(e) apertus-8b



(f) ministral-3-3b

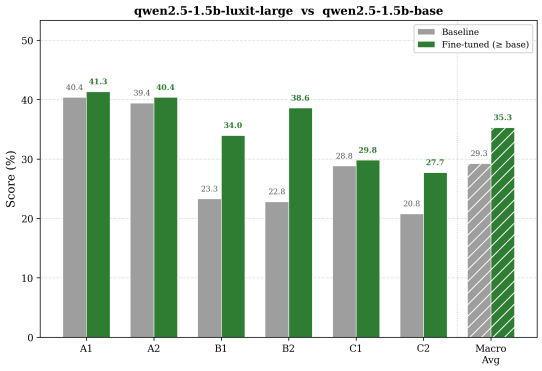


(g) llama-3.1-8b

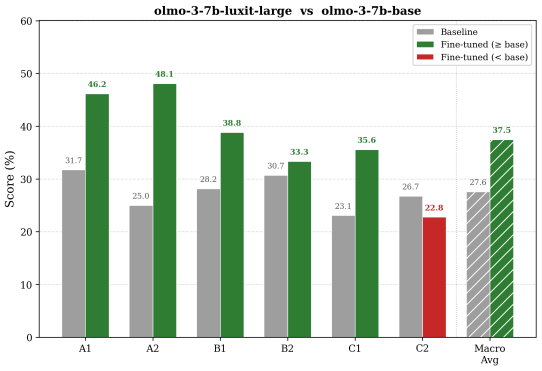


(h) mistral-7b

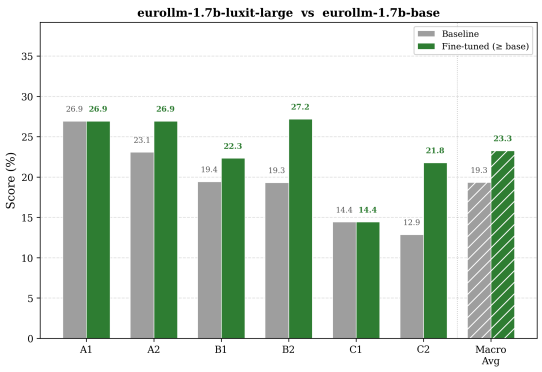
Figure 3: Total accuracy of fine-tuned models (labeled luxit-large) compared to their baselines on Luxembourgish language exams (A1–C2). Green bars indicate improvement over baseline; red bars indicate regression. Models are ordered roughly by macro-average performance of the fine-tuned variant (best to worst, left-to-right, top-to-bottom).



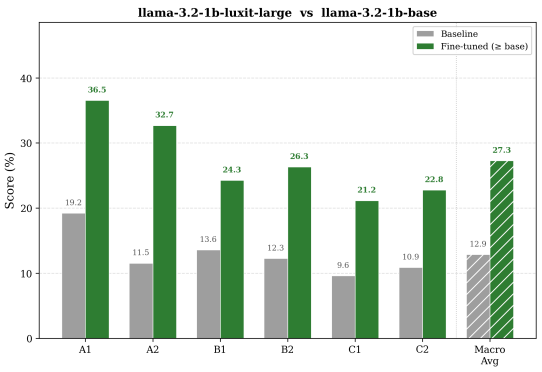
(i) qwen2.5-1.5b



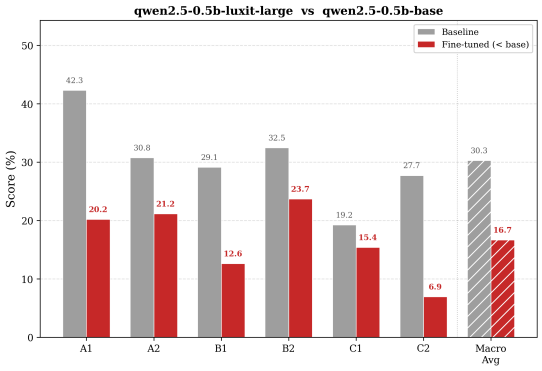
(j) olmo-3-7b



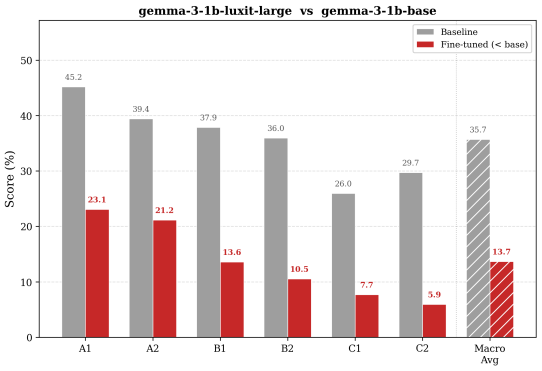
(k) eurollm-1.7b



(l) llama-3.2-1b



(m) qwen2.5-0.5b



(n) gemma-3-1b

Figure 3: (Continued) Total accuracy of fine-tuned models compared to their baselines on Luxembourgish language exams (A1–C2).

Score type	Score 1	Score 2	Score 3	Mean	Median
Original Dataset (245,624 entries)					
LQ	6,924 (2.8%)	189,858 (77.3%)	48,842 (19.9%)	2.171	2.0
FA	11,745 (4.8%)	101,693 (41.4%)	132,186 (53.8%)	2.490	3.0
IA	1,039 (0.4%)	21,520 (8.8%)	223,065 (90.8%)	2.904	3.0
HR	246 (0.1%)	44,755 (18.2%)	200,623 (81.7%)	2.816	3.0
Filtered Dataset (227,507 entries)					
LQ	N/A	180,075 (79.2%)	47,432 (20.8%)	2.208	2.0
FA	N/A	97,068 (42.7%)	130,439 (57.3%)	2.573	3.0
IA	N/A	16,343 (7.2%)	211,164 (92.8%)	2.928	3.0
HR	N/A	29,305 (12.9%)	198,202 (87.1%)	2.871	3.0

Table 6: LuxIT Score Distribution. We report scores from post-filtering on LuxIT with scores 1 (low), 2 (acceptable) and 3 (excellent) for Linguistic Quality (LQ), Factual Accuracy (FA), Instruction Adherence (IA) and Helpfulness Relevance (HR). We show the score distribution on the original dataset compared with the filtered dataset after rejecting 18,117 samples with low scores.

Evaluation Metric	Fleiss' κ	Aggregated Mean	% Acceptable (Score 2)	% Excellent (Score 3)
Linguistic Quality	0.243	2.33	66.7%	33.3%
Factual Accuracy	-0.019	2.94	5.8%	94.2%
Instruction Adherence	0.102	2.93	7.2%	92.8%
Helpfulness Relevance	0.090	2.80	20.3%	79.7%

Table 7: Human evaluation results on a random subset of 100 samples. Inter-annotator agreement is reported using Fleiss' κ across three independent evaluators. The final aggregated scores (Majority Vote) are reported for the 69 retained samples (scores of 1 "Low" are discarded).

Model	GPU	4-bit Load	BS	GAS	LR	Target Modules
Gemma-3-12B-IT	RTX 4090	Yes	2	16	1×10^{-4}	Gemma-specific[1]
Olmo-3-7B-Instruct	RTX 4090	No	8	4	2×10^{-4}	Standard[2]
EuroLLM-1.7B-Instruct	RTX 4090	No	16	3	1×10^{-4}	Standard[2]
Gemma-3-1B-IT	RTX 4090	No	16	3	1×10^{-4}	Gemma-specific[1]
Apertus-8B-Instruct[3]	RTX 4090	No	2	16	2×10^{-4}	all-linear
Qwen2.5-7B-Instruct	V100	No	2	16	1×10^{-4}	Standard[2]
Qwen2.5-1.5B-Instruct	V100	No	2	16	2×10^{-4}	Standard[2]
Qwen2.5-0.5B-Instruct	V100	No	2	16	2×10^{-4}	Standard[2]
Phi-4	V100	Yes	2	16	1×10^{-4}	Standard[2]
Mistral-7B-Instruct-v0.3	V100	No	4	8	1×10^{-4}	Standard[2]
Ministral-3-3B	V100	No	4	10	2×10^{-4}	Standard[2]
Llama-3.2-1B-Instruct	V100	No	8	6	2×10^{-4}	Standard[2]
Llama-3.1-8B-Instruct	V100	No	2	16	1×10^{-4}	Standard[2]
GLM-4-9B-0414	V100	No	2	16	1×10^{-4}	Standard[2]

Table 8: Hyperparameter Settings by Model. Constants across all models include 2 training epochs, AdamW 8-bit optimizer, cosine learning rate scheduler, a warmup ratio of 0.03, weight decay of 0.01, and LoRA parameters ($r = 16, \alpha = 32$, dropout = 0.05). BS = Train Batch Size per Device, GAS = Gradient Accumulation Steps, LR = Learning Rate. [1]Gemma-specific: vision_layers=False, language_layers=True, attention_modules=True, mlp_modules=True. [2]Standard: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj. [3]Apertus-8B-Instruct was the only model fine-tuned utilizing standard Hugging Face pipelines rather than Unsloth.

	IC			RTE			SA			SST-2			WNLI			
Base Model	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	Base	LuxIT	Δ	**Avg. Δ**
phi-4	0.647	0.695	+0.048	0.449	0.775	+0.326	0.017	0.430	+0.413	0.015	0.816	+0.802	0.293	0.327	+0.034	**+0.325**
olmo-3-7b	0.150	0.245	+0.096	0.142	0.666	+0.524	0.249	0.218	−0.031	0.300	0.787	+0.488	0.285	0.265	−0.020	**+0.211**
qwen2.5-7b	0.491	0.702	+0.210	0.658	0.774	+0.116	0.323	0.349	+0.026	0.285	0.770	+0.485	0.511	0.536	+0.025	**+0.172**
qwen2.5-0.5b	0.008	0.000	−0.008	0.276	0.392	+0.115	0.065	0.076	+0.011	0.297	0.504	+0.207	0.331	0.390	+0.059	**+0.077**
mistral-7b	0.412	0.419	+0.008	0.548	0.716	+0.168	0.354	0.383	+0.029	0.652	0.767	+0.115	0.265	0.290	+0.025	**+0.069**
gemma-3-1b	0.150	0.186	+0.036	0.530	0.560	+0.031	0.246	0.251	+0.005	0.636	0.694	+0.058	0.293	0.482	+0.189	**+0.064**
llama-3.1-8b	0.518	0.545	+0.027	0.332	0.623	+0.291	0.188	0.348	+0.160	0.711	0.651	−0.060	0.520	0.320	−0.200	**+0.044**
eurollm-1.7b	0.006	0.019	+0.013	0.139	0.054	−0.085	0.052	0.213	+0.161	0.267	0.289	+0.023	0.318	0.227	−0.090	**+0.004**
llama-3.2-1b	0.010	0.009	−0.001	0.251	0.324	+0.072	0.106	0.082	−0.024	0.175	0.338	+0.163	0.459	0.265	−0.194	**+0.003**
ministral-3-3b	0.398	0.322	−0.076	0.392	0.551	+0.159	0.425	0.192	−0.233	0.545	0.684	+0.139	0.329	0.324	−0.005	−0.003
qwen2.5-1.5b	0.232	0.142	−0.090	0.310	0.414	+0.105	0.263	0.240	−0.023	0.565	0.501	−0.064	0.385	0.390	+0.006	−0.013
glm-4-9b	0.684	0.743	+0.059	0.775	0.735	−0.040	0.307	0.328	+0.021	0.823	0.838	+0.016	0.478	0.311	−0.167	−0.022
gemma-3-12b	0.865	0.342	−0.522	0.744	0.770	+0.025	0.524	0.455	−0.068	0.783	0.716	−0.067	0.569	0.564	−0.005	−0.128
apertus-8b	0.576	0.460	−0.116	0.686	0.010	−0.677	0.505	0.234	−0.271	0.706	0.459	−0.247	0.265	0.543	+0.278	−0.206

Table 10: Complete evaluation results comparing **Baseline** and **LuxIT** fine-tuned F1-Macro scores across five Luxembourgish NLP tasks: Intent Classification (IC), Recognizing Textual Entailment (RTE), Sentiment Analysis (SA), Stanford Sentiment Treebank (SST-2), and Winograd Natural Language Inference (WNLI). Δ = LuxIT − Baseline; bold values in the *Avg.* Δ column indicate net improvement. Models are sorted by average Δ in descending order; the horizontal rule separates positive from negative average gains.

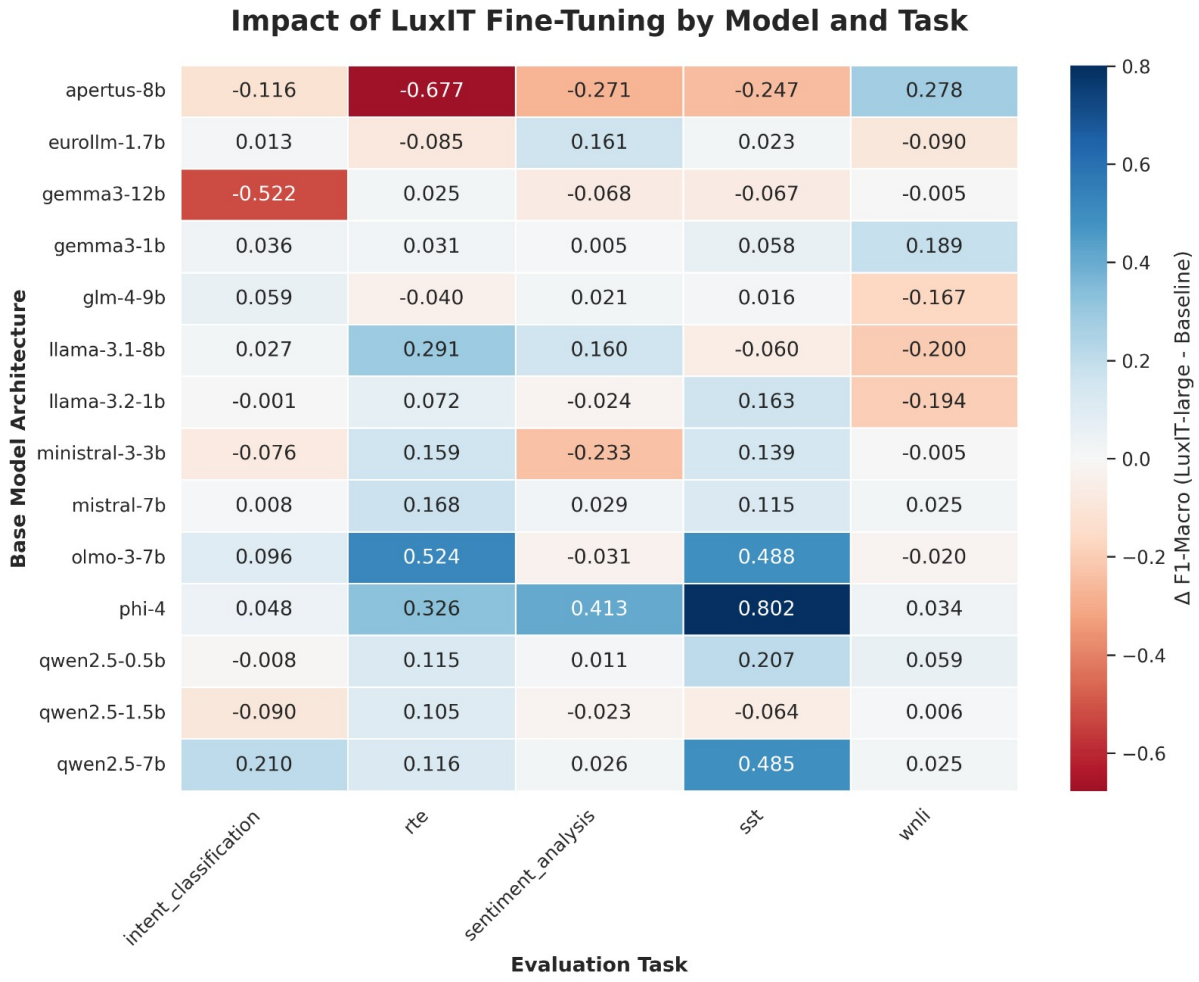


Figure 4: Impact of LuxIT Fine-Tuning by Model and Task (ΔF1-Macro).