LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data

Anonymous ACL submission

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 novel, 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., 2024; 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., 2024). 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., 2024). 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, low-resource language instruction tuning datasets, as existing IT datasets are predominantly in English, marginalizing other languages (Weber et al., 2024). This disparity leads to inferior performance for low-resource languages and increases the cost associated with their training and deployment (Weber et al., 2024).

This lack of resources is felt most acutely for low-resource languages, such as Luxembourgish, a West Germanic language spoken by about 600 000 people[1], with its primary concentration of speakers found in Luxembourg. In this paper, we introduce LuxIT, a monolingual instruction tuning dataset in Luxembourgish, 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

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

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

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, no instruction tuning dataset for Luxembourgish currently exists that has been created exclusively from monolingual seed data for the purpose of fine-tuning state-of-the-art LLMs.

Our primary contributions are twofold:

- We introduce LuxIT[3], a synthetically generated instruction tuning dataset in Luxembourgish derived from monolingual seed data.
- We evaluate the utility of LuxIT by fine-tuning 14 widely-used small-scale LLMs ($\leq$ 15B parameters) 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: Does fine-tuning on LuxIT lead to a general improvement in the Luxembourgish capabilities of LLMs? 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).

RQ2: To what extent does fine-tuning on LuxIT impact specific linguistic categories? We examine the results from RQ1 further and analyze them over the individual categorical exam levels, akin to Lothritz et al. (2026).

RQ3: Does fine-tuning on LuxIT lead to a general improvement in Luxembourgish NLP task performance? We benchmark our fine-tuned models against their respective base instruct models on 5 well-known NLP tasks in Luxembourgish.

2 Related Work

2.1 Multilingual Instruction Tuning Datasets

xP3 (Muennighoff et al., 2023) is a multilingual, human-crafted instruction tuning dataset, where the data is taken from P3 and other multilingual datasets (Zhang et al., 2024). The dataset is constructed by combining data from various sources into a unified format (Zhang et al., 2024). Muennighoff et al. (2023) further extend xP3 to xP3mt by applying machine translation (Weber et al., 2024).

Multilingual datasets like Bactrian-X (Li et al., 2023b) emerged from machine-translated Alpaca instructions and matching GPT-3.5-Turbo (Brown et al., 2020) generated answers (Weber et al., 2024). LIMA (Zhou et al., 2023) is a fine-tuned version of Llama (Touvron et al., 2023a), trained on 1000 attentively selected instruction-answer pairs (Zhang et al., 2024; Weber et al., 2024). Weber et al. (2024) extend LIMA into Lima-X by translating the instructions from the LIMA dataset into 4 languages.

We refrain from using machine translation and multilingual seed-data and investigate the dataset generation in a low-resource setting from a strictly monolingual approach. While the model we use for data generation is pre-trained on a multilingual corpus, the data we feed to the model for synthesizing the IT dataset is exclusively in 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 pre-training dataset in Luxembourgish through a data-augmentation technique where they partially translated text from German into Luxembourgish, resulting in LuxemBERT. Plum et al. (2024) 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. The cross-lingual dataset from Philippy et al. (2025a) demonstrates an effective strategy for generating instruction data, pairing Luxembourgish with English, German, or French content. However, this approach is contingent on the availability of parallel seed data, which is often a bottleneck for low-resource languages. Our work complements this by

[3]We provide a Wikipedia subset here and the code here

exploring a monolingual methodology specifically designed for scenarios where such parallel data is unavailable. This focus on a strictly monolingual setting is the primary distinction of our approach and motivates the investigation into creating resources under these more constrained conditions.

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.

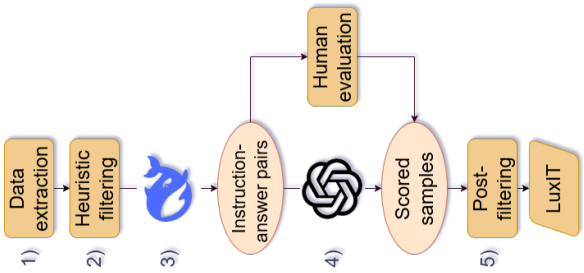


Figure 1: LuxIT Data Generation Pipeline.

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[5] 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. 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[6] 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[7].

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 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 three columns: `instruction`, `response`, and `messages`. The `instruction` and `response` columns store the

[4] https://dumps.wikimedia.org/lbwiki/latest/
[5] https://pypi.org/project/wikiextractor/0.1/
[6] While Lothritz et al. (2026) used the original Deep-Seek-R1 in their experiments, we use the latest DeepSeek-R1-0528
[7] There were some parsing errors for the generated json strings, hence we exclude those samples.

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 3 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 4 (Appendix A.5) which shows that despite a slight underperformance at lower 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. 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

Using the same scoring rubric, we manually evaluate a small subset of 100 generated samples, with their scores aggregated via majority vote across 3 Luxembourgish native speakers. 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 (κ = -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 full results in Table 6 in Appendix A.8.1.

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 5 and show the score distribution on LuxIT from our post-filtering step. Due to space reasons, we put the table in Appendix A.8. 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.

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 209307 and evaluating on 11375 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 (Apertus et al., 2025), EuroLLM-1.7B-Instruct (Martins et al., 2024), Olmo-3-7B-Instruct (Olmo et al., 2025). The models are fine-tuned with LoRA (Hu et al., 2021) using the Unsloth[8] 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 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)*[9] and are divided into 6 skill levels defined in the Common European Framework of Reference for Languages (Council of Europe, 2001), 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 base and instruct models to our fine-tuned models.

The INLL exams are proprietary and not publicly accessible. This effectively eliminates 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.

[8] https://docs.unsloth.ai/
[9] https://www.inll.lu/en/

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), Winograd Natural Language Inference (WNLI), Recognizing Textual Entailment (RTE), Sentiment Analysis and Stanford Sentiment Treebank (SST-2), all 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 8 in Appendix B.2.

5 Results

5.1 RQ1: Does fine-tuning on LuxIT lead to a general improvement in the Luxembourgish capabilities of LLMs?

Table 1 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-Instruct) 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-Instruct 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-Instruct 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

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-0414	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 1: 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).

contains a more detailed visual breakdown.

5.2 RQ2: To what extent does fine-tuning on LuxIT impact specific linguistic categories?

Table 2 shows macro-averaged accuracy changes across the four skill categories introduced in section 4.2. The most consistent 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). Reading 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 1.

5.3 RQ3: Does fine-tuning on LuxIT lead to a general improvement in Luxembourgish NLP task performance?

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, 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.

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 2: 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).

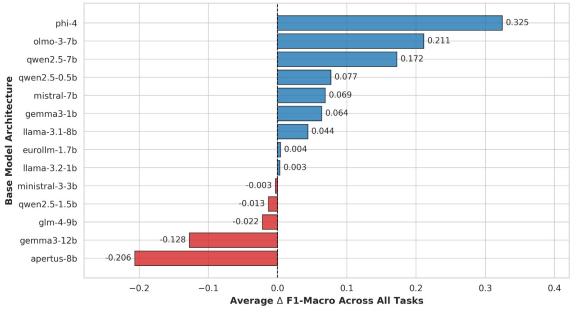


Figure 2: Average Impact of LuxIT Fine-Tuning per Model across all tasks.

6 Discussion

LuxIT represents an important step toward addressing the scarcity of IT resources for Luxembourgish. Leveraging DeepSeek-R1-0528 for its demonstrated Luxembourgish proficiency, our pipeline effectively synthesizes a large dataset from monolingual seed data. While the discrepancy in retention rates (92.6% LLM vs. 69.0% human) highlights a leniency bias in the automated judge, majority-vote aggregation over the retained samples points toward generally high factual and instructional quality, while also highlighting the need for more objective evaluation frameworks. The low inter-annotator agreement, including a negative Fleiss' κ for factual accuracy, further underscores the inherent subjectivity of evaluating synthetic Luxembourgish text. Despite this subjectivity, the human evaluation yields a higher proportion of "Excellent" linguistic quality scores (33.3%) than the LLM (20.8%), suggesting that the LLM judge may be more critical of stylistic nuances than humans.

Our fine-tuning experiments show highly promising results for RQ1: 12 of 14 models improve in macro-average accuracy, with gains ranging from +2.5 pp (Gemma-3-12B-Instruct) to +21.0 pp (Ministral-3-3B-Instruct). The divergent outcomes between Llama-3.2-1B-Instruct (+14.4 pp) and Gemma-3-1B-Instruct (-22.0 pp) suggest that fine-tuning receptivity depends on more than model size alone. Llama-3.2-1B's very low baseline (12.9%) indicates limited prior task alignment, which LuxIT appears to correct effectively. The degradations observed for Gemma-3-1B-Instruct and Qwen2.5-0.5B-Instruct, both with relatively stronger baselines (35.7% and 30.3% respectively), may reflect catastrophic forgetting of pre-trained representations, though differences in architecture and model capacity at sub-1.5B scale are equally plausible contributing factors. Disentangling these effects would require controlled ablations beyond the scope of this work.

A deeper analysis of the language exam levels reveals that gains from LuxIT are not uniformly distributed. Most models exhibited their most consistent improvements in the intermediate B1 and B2 levels. Contrarily, the advanced C2 level saw the most regression, with 6 out of 14 models declining in accuracy. One possible explanation is that the synthetic data generated by DeepSeek-R1-

0528 is skewed towards B1 and B2 proficiency levels. It is possible that the generator model itself lacks the capacity to consistently produce the nuanced expressions required at a native-like C2 level in Luxembourgish. However, further qualitative studies on the dataset's linguistic distribution are needed to substantiate this hypothesis.

Analyzing the categorical evaluation results further confirms that fine-tuning on LuxIT generally improves Luxembourgish language understanding, with the majority of ΔBase values being positive across all four skill categories. These categorical improvements closely parallel the general accuracy trends discussed previously. Specifically, the uniform performance drops seen in two of the sub-1.5B scale models (Qwen2.5-0.5B-Instruct and Gemma-3-1B-IT) across all four skills mirror their overall struggles. Contrarily, models in the 7–14B parameter range consistently show positive ΔBase values, with only marginal degradations. We observe the largest gains by fine-tuning on LuxIT in the vocabulary and conversational comprehension categories, with Llama-3.2-1B-Instruct and Phi-4 exhibiting remarkable improvements in vocabulary accuracy by +24.4 pp and +18.3 pp, respectively, suggesting that the inclusion of diverse Wikipedia and RTL news content effectively expands the models' Luxembourgish lexicon. These findings indicate that synthetic data can help build a model's vocabulary and comprehension skills for Luxembourgish, even if broader grammatical improvements remain more difficult to achieve.

The limited agreement between language exam results and NLP task performance, highlighted in RQ3, suggests that fine-tuning on LuxIT improves two partially distinct capabilities that are not interchangeable as evaluation targets. Beyond the difference in task format, multiple-choice comprehension questions versus structured label prediction, the two benchmarks differ fundamentally in the nature of the language they contain. The language exams consist of standardized, well-formed Luxembourgish produced by professional linguists at the INLL, whereas the NLP benchmark tasks contain naturally occurring, noisy, and non-standardized Luxembourgish that does not conform to the same linguistic norms. Since LuxIT is synthesized from Wikipedia and RTL news articles, the fine-tuned models are exposed to relatively formal and topically constrained Luxembourgish, which plausibly aligns better with the register of the language exams than with the heterogeneous language of the downstream tasks. The divergent results for individual models further support this interpretation: models such as ministral-3-3b appear to gain substantially in fluency and comprehension without acquiring the ability to produce constrained output labels reliably, while smaller models such as gemma-3-1b show the inverse pattern, suggesting that format alignment and linguistic improvement are separable effects of fine-tuning that LuxIT does not uniformly induce across architectures.

In terms of absolute capability, the results demonstrate that fine-tuning medium-sized models (~9B to 12B parameters) on LuxIT yields the most proficient Luxembourgish models. GLM-4-9B-0414 emerged as the top-performing model overall with a final macro average of 60.6%, closely followed by Gemma-3-12B-Instruct at 59.4%.

7 Conclusion

This work addresses the scarcity of instruction tuning resources for low-resource languages by introducing LuxIT, a high-quality dataset of 227,507 instruction-answer pairs in Luxembourgish synthesized exclusively from monolingual seed texts using DeepSeek-R1-0528. Automated filtering retains 92.6% of generated pairs, while a small-scale human evaluation highlights a leniency bias in the automated judge and the inherent subjectivity of assessing synthetic Luxembourgish text. Fine-tuning 14 small-scale LLMs on LuxIT yields improvements in macro-average language exam accuracy for 12 of 14 models (+5.37 pp average), with the strongest gains at intermediate B1/B2 proficiency levels and in vocabulary and conversational comprehension. On five Luxembourgish NLP downstream tasks, 9 of 14 models improve in macro-averaged F1, with the largest gains for phi-4 (+0.325), olmo-3-7b (+0.211), and qwen2.5-7b (+0.172). However, language exam and NLP task improvements are not systematically correlated, suggesting LuxIT improves formal comprehension and fluency more reliably than structured label prediction on naturally occurring text. Results vary among the smallest architectures (<1.5B parameters), highlighting a tension between task alignment and catastrophic forgetting. Overall, our findings validate monolingual synthetic data generation as a feasible strategy for enhancing LLM capabilities in low-resource language settings. Future work will explore open-ended conversational evaluations and multi-turn data integration.

8 Limitations

We acknowledge several limitations. First, the source data for LuxIT is originating from Wikipedia and RTL news articles. This domain specificity means the dataset may lack conversational, dialectal, or creative language styles.

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). While using a single model was a necessary computational trade-off, it suggests the full dataset contains marginally more noise than the human-verified subset. Furthermore, the low inter-annotator agreement (e.g., negative Fleiss' Kappa for factual accuracy) in the human audit indicates our guidelines allowed for highly subjective interpretations among native speakers, emphasizing 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.

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.

Fifth, our downstream evaluation is restricted to multiple-choice language proficiency exams. 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.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Council of Europe. 2001. Common European framework of reference for languages: Learning, teaching, assessment.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 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, 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 Arun Rao et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego

de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 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, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Cedric Lothritz, Jordi Cabot, and Laura Bernardy. 2026. Testing low-resource language support in llms using language proficiency exams: the case of luxembourgish. *Preprint*, arXiv:2504.01667.

Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé Bissyandé, Jacques Klein, Isabella Olariu, Andrey Boytsov, Clément LeFebvre, and Anne Goujon. 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, 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.

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

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

Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. *Preprint*, arXiv:2402.18334.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 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.

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. 2024. Text generation models for luxembourgish with limited data: A balanced multilingual strategy. *Preprint*, arXiv:2412.09415.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, and 22 others. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

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.

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 Gaël Liu et al. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and Brian Fuller 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, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? *Preprint*, arXiv:2402.13703.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, LEI BAI, Jing Shao, and Wanli Ouyang. 2023. Lamm: Language-assisted multimodal 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, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

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[10] 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 title>',
 '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:

[10] We do not show title, header and text for the RTL news article as the data is not publicly available

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",
       "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 3: 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 3). 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 4 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 4: 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 5 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 6.

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 7. (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 8 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)	2360
Winograd NLI (WNLI)	136

Table 8: Sample distribution across Luxembourgish downstream tasks.

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 5: 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 6: 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).

C Results

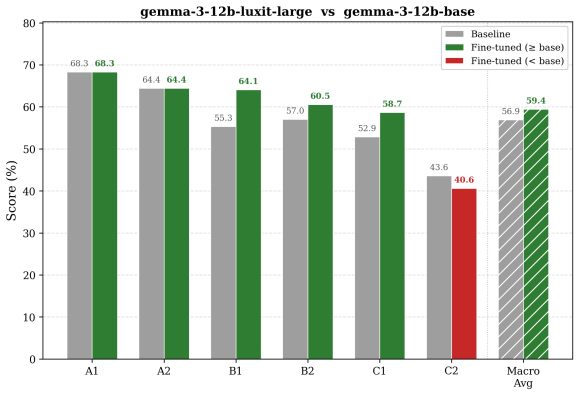
Here we provide more details accompanying the Results section (Section 5).

C.1 RQ1

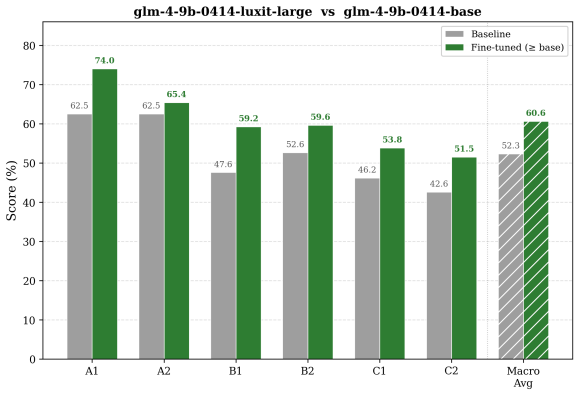
We provide detailed visual comparison between our fine-tuned models and their respective baseline in Figure 3 (Section 5.1).

C.2 RQ3

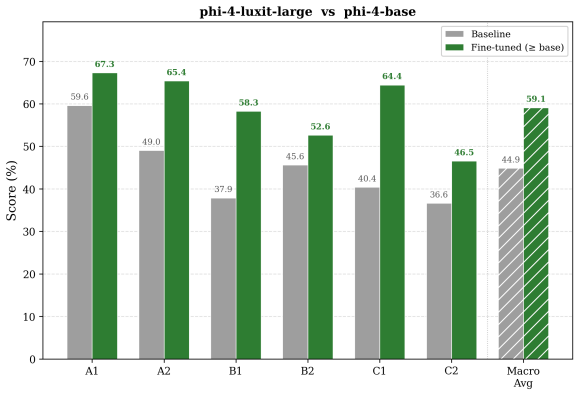
We report the results of benchmarking our fine-tuned models against their respective baseline models (Section 5.3) on the Luxembourgish downstream tasks in Table 9 (introduced in Section 4.3) and show the per-task changes and the average change per model in Figure 4 and Figure 2 respectively.



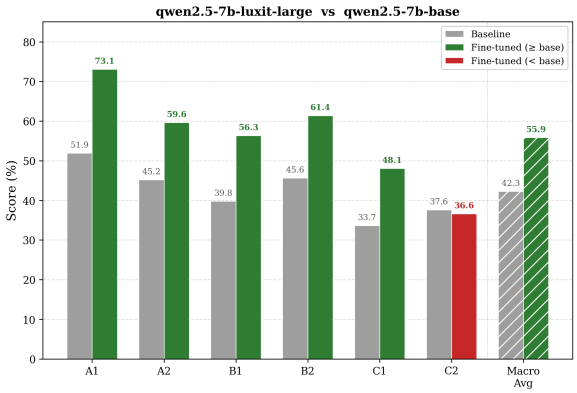
(a) gemma-3-12b



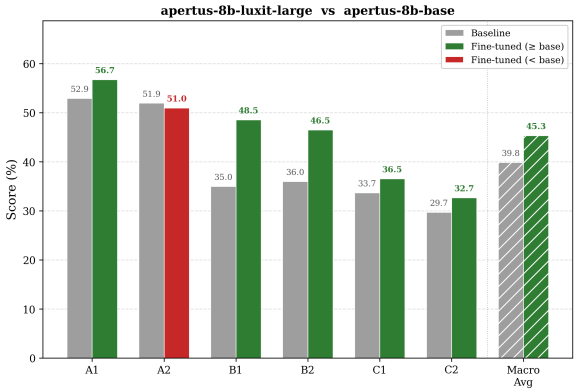
(b) glm-4-9b-0414



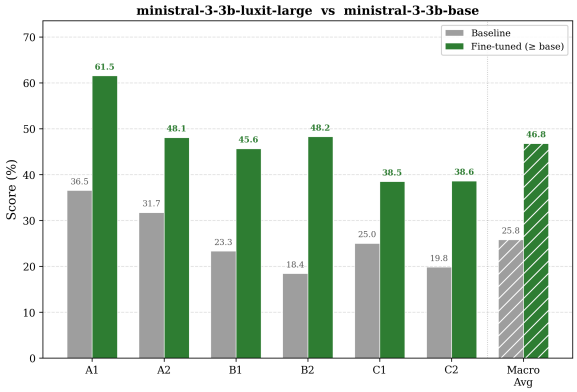
(c) phi-4



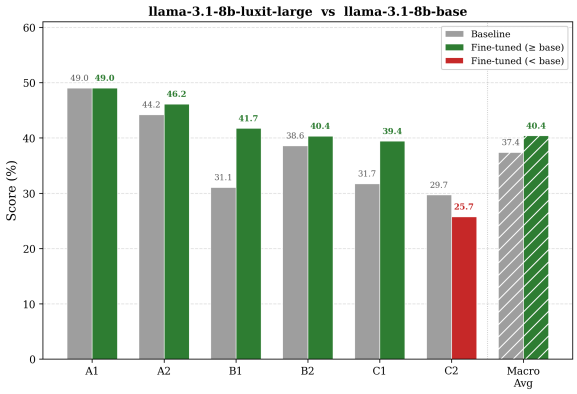
(d) qwen2.5-7b



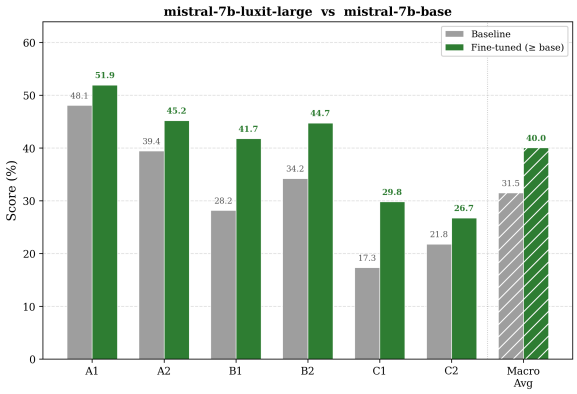
(e) apertus-8b



(f) ministral-3-3b

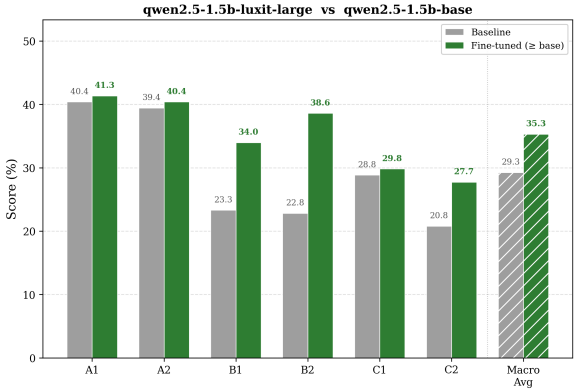


(g) llama-3.1-8b

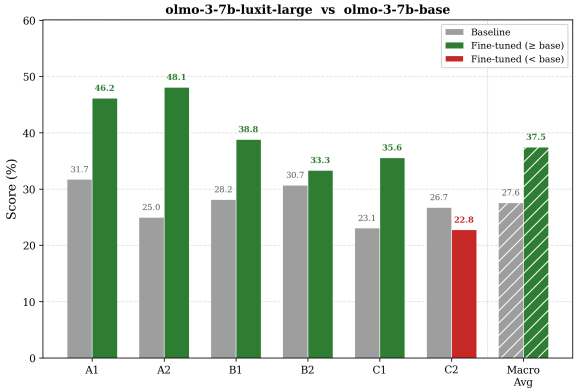


(h) mistral-7b

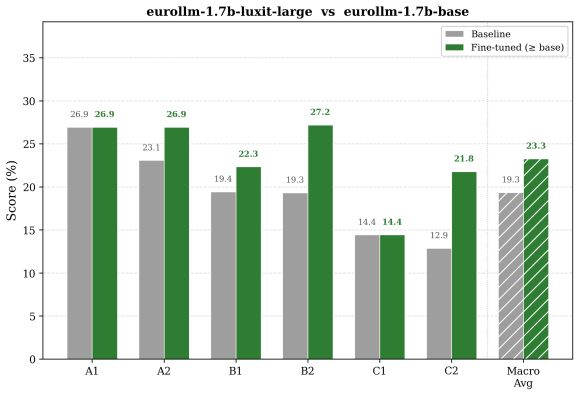
Figure 3: Total accuracy of fine-tuned models compared to their baselines on Luxembourgish language exams (A1–C2). Green bars indicate improvement over baseline; red bars indicate regression. Models are ordered roughly by macro-average performance of the fine-tuned variant (best to worst, left-to-right,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



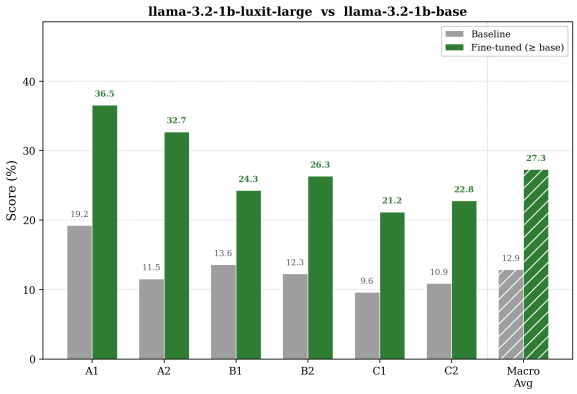
(i) qwen2.5-1.5b



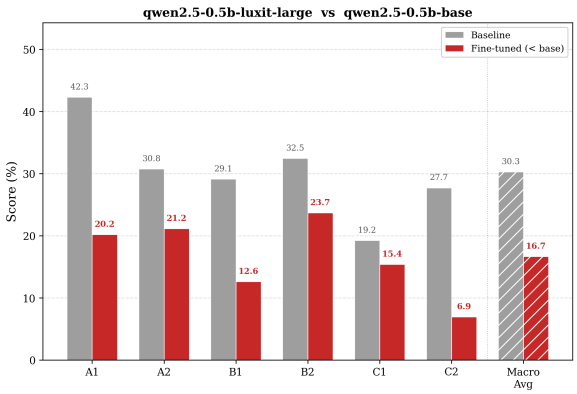
(j) olmo-3-7b



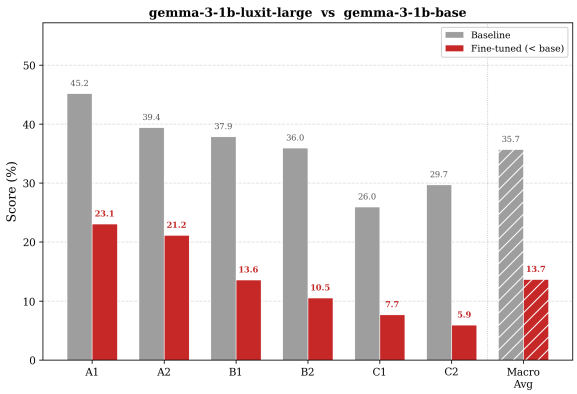
(k) eurollm-1.7b



(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).

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 7: 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			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 9: Complete evaluation results comparing **Baseline** and **LuxIT-Large** fine-tuned F1-Macro scores across five Luxembourgish NLP tasks: Intent Classification (IC), Recognising Textual Entailment (RTE), Sentiment Analysis (SA), Sentiment Strength Tagging (SST), 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.

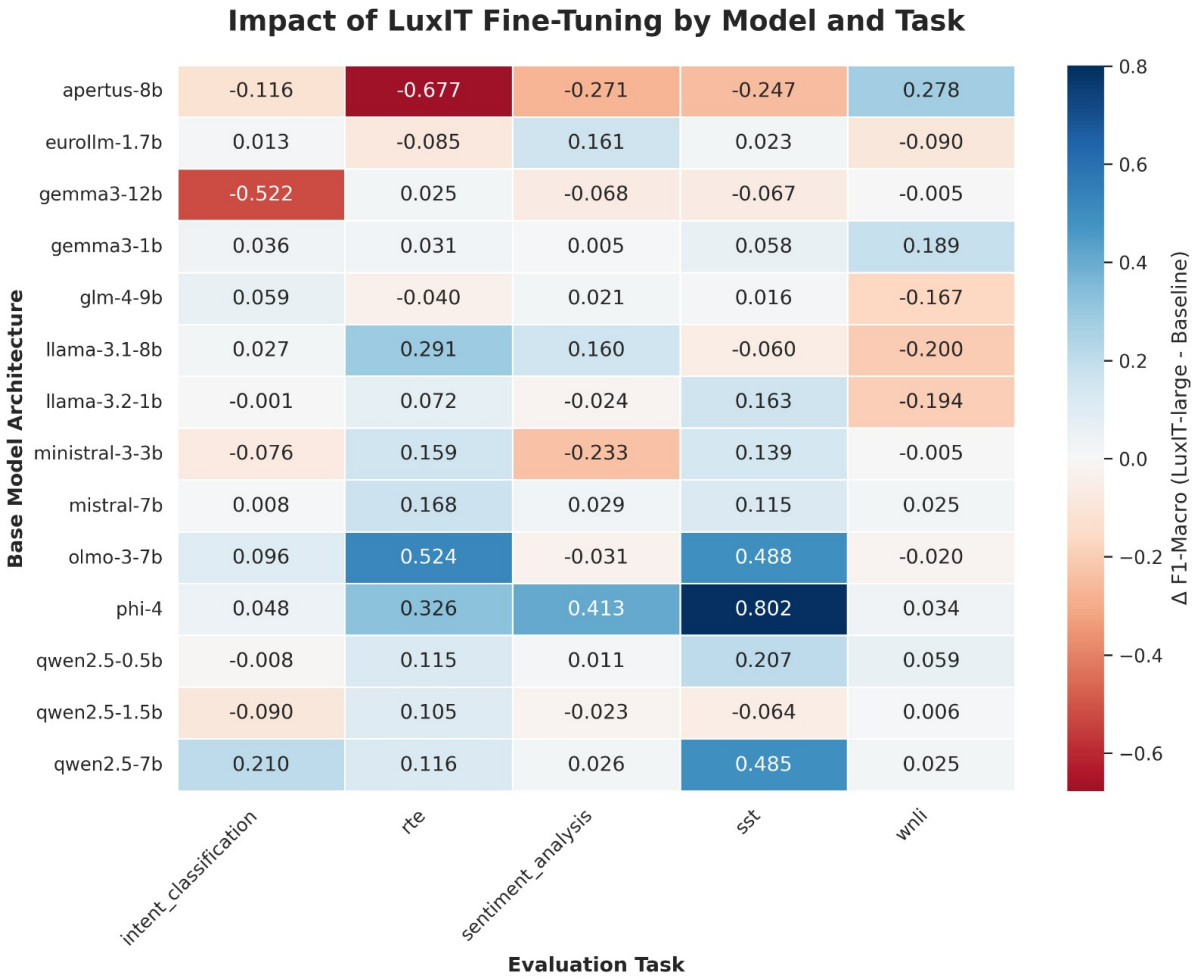


Figure 4: Impact of LuxIT Fine-Tuning by Model and Task (ΔF1-Macro).