

# LuxIT: A Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data

Julian Valline, Cédric Lothritz, Jordi Cabot

Luxembourg Institute of Science and Technology  
5 Av. des Hauts-Fourneaux L-4362 Esch-sur-Alzette  
{julian.valline, cedric.lothritz, jordi.cabot}@list.lu

## 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. To investigate the practical utility of the dataset, we fine-tune several smaller-scale LLMs on LuxIT. Subsequent benchmarking against their base models on Luxembourgish language proficiency examinations, however, yields mixed results, with performance varying significantly across different models. LuxIT represents a critical contribution to Luxembourgish natural language processing and offers a replicable monolingual methodology, though our findings highlight the need for further research to optimize its application.

**Keywords:** instruction-tuning, data augmentation, low-resource languages

## 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 like 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, non-English instruction tuning datasets, as the majority of existing IT datasets

are predominantly in English, marginalizing other languages (Weber et al., 2024). This disparity leads to inferior performance for non-English 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<sup>1</sup>, 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<sup>2</sup> and Wikipedia entries, both in Luxembourgish. Following the work of Lothritz and Cabot (2025), we select DeepSeek-R1-0528, as the top-performing model for Luxembourgish generation and comprehension to synthetically produce 59 242 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:

<sup>1</sup><https://cursus.edu/en/23040/luxembourgish-at-its-best>

<sup>2</sup><https://www.rtl.lu/>

- We introduce LuxIT<sup>3</sup>, a synthetically generated instruction tuning dataset in Luxembourgish derived from monolingual seed data.
- We examine the extent to which fine-tuning on LuxIT changes the Luxembourgish capabilities of 5 widely-used LLMs.

## 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 is combined 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 original 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 instruction tuning dataset is only in Luxembourgish.

### 2.2. Synthesizing Instruction Tuning Data through Knowledge Distillation

Taori et al. (2023) developed Alpaca, a modified Llama-7B model fine-tuned on 52k instruction-input-response triplets (Zhang et al., 2024), which are generated with text-davinci-003 (Brown et al., 2020). Alpaca performed on par with text-davinci-003 from OpenAI (Zhang et al., 2024).

Peng et al. (2023) use GPT-4 (OpenAI et al., 2024b) to generate the output for the instruction-input pairs used in the original Alpaca dataset.

Our work follows a similar dataset structure as Peng et al. (2023), except that we use DeepSeek-

R1-0528 to generate all new instructions and expected answers in Luxembourgish.

### 2.3. 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 LuxeBERT. 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 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 involves the extraction of raw data from two distinct sources (1) (see Section 3.1), which we then subject to a series of heuristic filters (2) (see Section 3.1.1). 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. We choose GPT-5-mini (OpenAI, 2025a) for the judging task, primarily because of its favorable cost-

<sup>3</sup>We provide a large subset sourced only from Wikipedia articles [here](#)

performance trade-off. Lothritz and Cabot (2025) place GPT-4o-mini (OpenAI et al., 2024a) among the top-performing models for Luxembourgish, although it is significantly outperformed by the larger GPT-4o model. We assume a similar performance relationship holds between GPT-5 and GPT-5-mini, and thus utilize the highly cost-effective GPT-5-mini, expecting it to at least match or exceed the previous generation’s performance. Additionally, we perform a small-scale pilot study where we compare the judging of GPT-5 and GPT-5-mini, finding that the retention rate is only off by a few percent, further justifying this cost-performance trade-off. For the judging task, GPT-5-mini assigns scores to each pair based on predefined criteria (see Section 3.4). In parallel, a subset of these pairs undergoes manual human evaluation (4). In the final step, we apply a post-filtering process to remove all samples that received poor scores, resulting in the final LuxIT dataset (5).

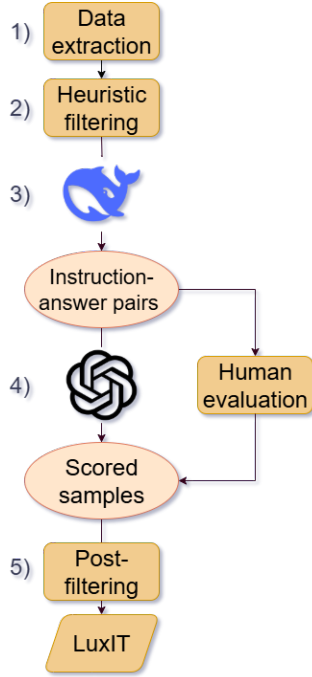


Figure 1: LuxIT Data Generation Pipeline.

### 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<sup>4</sup> on August 5th, 2025 and subsequently use Wikiextrac-

<sup>4</sup><https://dumps.wikimedia.org/lbwiki/latest/>

tor<sup>5</sup> 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.

#### 3.1.1. Filtering

Before the generation phase, we implement a set of heuristic filters to remove low-quality samples. We retain only articles containing a minimum of 750 characters. For Wikipedia articles, we specifically exclude list articles, disambiguation pages, and any pages still marked as stubs. For RTL news articles, we verify that each article is written in Luxembourgish 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.

### 3.2. Instruction-Answer pair generation

We employ DeepSeek-R1-0528<sup>6</sup> to synthetically create instruction-answer pairs in Luxembourgish from the filtered RTL news and Wikipedia articles. Due to resource limitations, we restrict the seed data to a randomly sampled subset of 9000 Wikipedia articles and 13 390 RTL news articles. We then prompt the model to generate three instruction-answer pairs for each seed, yielding a total of 66 005 synthetic pairs<sup>7</sup>.

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. 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.

<sup>5</sup><https://pypi.org/project/wikiextractor/0.1/>

<sup>6</sup>While Lothritz and Cabot (2025) used the original Deep-Seek-R1 in their experiments, we use the latest version DeepSeek-R1-0528

<sup>7</sup>There were some parsing errors for the generated json strings, hence we exclude those samples

### 3.3. Instruction Tuning Dataset

The final LuxIT dataset is structured with three columns: `instruction`, `response`, and `conversations`. The `instruction` and `response` columns store the generated instructions and their corresponding answers. The `conversations` column consolidates these two into the ShareGPT format. We show the structure of our prompt template in Appendix A.3 and the LuxIT structure in Table 1.

### 3.4. Post-filtering

Following the generation stage, we adopt an LLM-as-a-judge methodology for quality control, utilizing GPT-5-mini. We prompt the model to assess each instruction-answer 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 full breakdown of our scoring system is as follows:

#### *Linguistic Quality 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.
- 2: Mostly correct Luxembourgish, but has minor errors or sounds slightly robotic/unnatural. May mix in too many loan words unnecessarily.
- 3: Fluent, idiomatic, and grammatically perfect Luxembourgish. Natural-sounding text that a native speaker would produce.

#### *Factual Accuracy score*

- 1: Contains factual errors that contradict the source text or general knowledge.
- 2: Mostly accurate but might have minor inaccuracies or omissions.
- 3: Completely accurate according to the source text and factual knowledge.

#### *Instruction Adherence score*

- 1: Fails to follow the core instruction (e.g., provides a summary when asked for a list).
- 2: Follows the main instruction but misses a constraint (e.g., writes 4 bullet points when asked for 3, wrong format, or incorrect tone).
- 3: Perfectly follows all parts of the instruction, including constraints like length, format, and tone.

#### *Helpfulness Relevance score*

- 1: The instruction is nonsensical, irrelevant to any reasonable context, or the response is unhelpful/off-topic.
- 2: The instruction is plausible but not very insightful or creative. The response addresses the instruction but in a basic way.
- 3: A genuinely useful, interesting, or creative instruction that elicits a helpful, comprehensive response.

We provide the specific prompt used for this evaluation in Appendix A. We retain only samples that achieved 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. The evaluation is done by one of the authors, who is a Luxembourgish native speaker. We then compare the results of this manual assessment against the scores assigned by the LLM-as-a-judge (see Section 5.1.1).

#### 3.4.2. Statistics

The final version of our dataset contains 59,242 high-quality instruction-answer pairs in Luxembourgish.

## 4. Experimental Setup

In this section, we describe our experiments and outline the central research questions that this study addresses.

### 4.1. Research Questions

This study aims to address the following two research questions:

**RQ1: To what extent can a large language model with sufficient proficiency in Luxembourgish generate high-quality instruction-answer pairs in the language?** We address this question by executing our data generation pipeline (Figure 1) and assessing the generated outputs against four custom metrics designed to evaluate linguistic quality (see Section 3.4), thereby quantifying the proportion of high-quality samples.

**RQ2: Does fine-tuning on LuxIT lead to a general improvement in the Luxembourgish capabilities of LLMs?** To answer this, we fine-tune 5 LLMs with less than 10 billion parameters



Instruction	Response	Conversations
Wéi huert d'Beruffskarriär vum Aloyse Meyer bei der Diddelenger Schmelz an duerno bei der Arbed ausgesinn? Nëmmen d'Karriärsstationne bis 1920 bezéien.	Den Aloyse Meyer huert 1903 seng Karriär bei der Diddelenger Schmelz am Konstruktionsbüro ugefaangen. Kuerz duerno, nëmme e puer Méint drop, gouf hie 'Ingenieur adjoint'. 1906 krut hie d'Plaz als 'Chef de service'. 1912 gouf hien Direkter vun der Schmelzt. 1918 ass hien dann als techneschen Direkter an d'Zentraladministratioun vun der Arbed gewiesselt. Do gouf hien 1920 zum Generaldirekter ernannt.	[[{'from': 'human', 'value': "Wéi huert d'Beruffskarriär vum Aloyse Meyer bei der Diddelenger Schmelz an duerno bei der Arbed ausgesinn? Nëmmen d'Karriärsstationne bis 1920 bezéien."}, {'from': 'gpt', 'value': "Den Aloyse Meyer huert 1903 seng Karriär bei der Diddelenger Schmelz am Konstruktionsbüro ugefaangen. Kuerz duerno, nëmme e puer Méint drop, gouf hie 'Ingenieur adjoint'. 1906 krut hie d'Plaz als 'Chef de service'. 1912 gouf hien Direkter vun der Schmelzt. 1918 ass hien dann als techneschen Direkter an d'Zentraladministratioun vun der Arbed gewiesselt. Do gouf hien 1920 zum Generaldirekter ernannt."}]]

Table 1: LuxIT structure showing one example from the Wikipedia subset. We omit the `source_type` column

on LuxIT and benchmark their performance in Luxembourgish against their corresponding base instruct models, following the evaluation framework established by Lothritz and Cabot (2025).

## 4.2. Fine-tuning on LuxIT

We fine-tune a variety of small models (<10B parameters) on LuxIT. We select the following models: Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct (Grattafiori et al., 2024), Gemma-3-1B-IT (Team et al., 2025), GLM-4-9B-0414 (GLM et al., 2024) and Qwen-3-0.6B (Yang et al., 2025). The models are fine-tuned with LoRA (Hu et al., 2021) using the Unsloth<sup>8</sup> framework. We apply the following hyperparameters (Table 2):

Model	r	$\alpha$	Batch	LR
Llama 3.1 (8B)	16	32	32	2e-4
Llama 3.2 (1B)	32	64	48	2e-4
Gemma 3 (1B)	64	64	16	1e-4
GLM 4 (9B)	16	32	32	1e-4
Qwen 3 (0.6B)	32	32	16	3e-4

Table 2: Model-training Hyperparameter Settings. We show LoRA rank, LoRA  $\alpha$ , **Batch** size and **Learning Rate** for each model.

We train each model on 16-bit precision with cosine scheduler, AdamW\_8bit optimizer, warmup

<sup>8</sup><https://docs.unsloth.ai/>

ratio = 0.03, weight decay = 0.01 and a max sequence length of 2048 for a total of 3 epochs on a Tesla V100 with 32GB. We apply LoRA dropout = 0 for 0.6B & 1B models and LoRA dropout = 0.05 for 8B & 9B models.

## 4.3. Evaluation on Language Exams

We investigate the impact of LuxIT on the Luxembourgish proficiency of small LLMs. Specifically, we evaluate their language fluency, following the same approach proposed by Lothritz and Cabot (2025). 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 listening comprehension (done through transcripts). The language exams stem from the Luxembourgish language institute *Institut National des Langues Luxembourg (INLL)*<sup>9</sup> and are divided into 6 levels, ranging from A1 (basic level) to C2 (native level), each exam level consisting of fill-in-the-blank and multiple choice questions (Lothritz and Cabot, 2025). 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, listening comprehension). We measure the models' performance using accuracy as a metric, akin to Lothritz and Cabot (2025). We compare base and instruct models to our fine-tuned models<sup>10</sup>.

<sup>9</sup><https://www.inll.lu/en/>

<sup>10</sup>Since the authors did not evaluate Qwen-3-base, we omit comparison to the base version. To the best of our

## 5. Results

### 5.1. RQ1: To what extent can a large language model with sufficient proficiency in Luxembourgish generate high-quality instruction-answer pairs in the language?

Table 3 shows the score distribution on LuxIT from our post-filtering step (see Figure 1).

The original dataset consists of 66 005 instruction-answer pairs. We reject 6763 samples with low scores, observing a retention rate of 89.8%. Among all samples, 8679 (13.1%) have all perfect scores (score 3) across all score types. We observe that `instruction adherence` scores best among all metrics, with 58,235 samples (88.2%) obtaining a score of 3. Furthermore, among the rejected samples, `factual accuracy` has the highest rejection rate, with 3071 entries having a score of 1. For the rejection rate, we observe 1772, 3071, 288 and 80 entries of score 1 (lowest) for linguistic quality, factual accuracy, instruction adherence and helpfulness relevance respectively.

#### 5.1.1. Human evaluation study

We perform a small-scale human evaluation study on the generated instruction-answer pairs. Out of the 100 randomly selected samples, we deem 90% to be of high-quality. 30 instruction-answer pairs have an all perfect score (score 3) across all score types. Here `linguistic quality` is the most frequent reason for rejection, accounting for 7 entries, with factual accuracy, instruction adherence and helpfulness relevance accounting for 1, 2 and 0 entries of score 1 (lowest) respectively. This suggests that GPT-5-mini might not capture linguistic imperfections for Luxembourgish on a high level.

### 5.2. RQ2: Does fine-tuning on LuxIT lead to a general improvement in the Luxembourgish capabilities of LLMs?

Table 4 shows the comparison between our fine-tuned models, instruct models and base models evaluated on Luxembourgish language exams, reporting total accuracy of all categories. The results are mixed. We observe that fine-tuning on LuxIT improves general accuracy on language exams for GLM-4-9b-0414 except for B2 and C1, where it

knowledge, GLM et al. (2024) did not release a base version for GLM-4-9b-0414.

worsens by about 2 percentage points for both and remains unchanged for A1. For Llama-3.1-8b, we only observe improvements by about 5 percentage points for B2 and C2. Accuracy drops by about 1 percentage point on B1 and becomes significantly worse for all other levels. For Qwen-3-0,6b, accuracy improves by almost 10 percentage points on A1, about 1 percentage point for levels A2 and C2, but drops significantly in accuracy for the other levels. The fine-tuned versions of Llama-3.2-1b and Gemma-3-1b scored poorly on the language exams compared to their instruct and base versions.

Model	A1	A2	B1	B2	C1	C2
Base Models						
Llama 3.1	<b>56.7</b>	33.7	<b>33.0</b>	33.3	30.8	<b>33.7</b>
Llama 3.2	26.0	19.2	<b>28.2</b>	21.1	<b>19.2</b>	12.9
Gemma 3	45.2	<b>36.5</b>	<b>38.8</b>	<b>40.4</b>	25.0	<b>33.7</b>
Instruct Models						
Llama 3.1	55.8	<b>42.3</b>	31.1	33.3	<b>44.2</b>	28.7
Llama 3.2	<b>26.9</b>	<b>30.8</b>	<b>28.2</b>	<b>22.8</b>	16.3	<b>20.8</b>
Gemma 3	<b>46.2</b>	33.7	32.0	39.5	<b>29.8</b>	<b>33.7</b>
GLM 4	<b>60.6</b>	53.8	47.6	<b>47.4</b>	<b>45.2</b>	39.6
Qwen 3	34.6	36.5	<b>35.9</b>	<b>41.2</b>	<b>25.0</b>	28.7
Fine-tuned Models						
Llama 3.1	33.7	30.8	30.1	<b>38.6</b>	33.7	<b>33.7</b>
Llama 3.2	17.3	04.8	08.7	06.1	03.8	09.9
Gemma 3	04.8	04.8	05.8	03.5	01.0	05.0
GLM 4	<b>60.6</b>	<b>57.7</b>	<b>50.5</b>	45.6	43.3	<b>47.5</b>
Qwen 3	<b>43.3</b>	<b>37.5</b>	23.3	22.8	20.2	<b>29.7</b>

Table 4: Results on Luxembourgish language exams. Best results are highlighted in **bold**.

## 6. Discussion

We think LuxIT is essential to the continued growth of Luxembourgish language resources. The methodology, centered on leveraging a state-of-the-art model (DeepSeek-R1-0528) selected for its existing Luxembourgish proficiency, proved effective in generating a substantial dataset from monolingual seed data. The subsequent quality con-

Score type	Score 1	Score 2	Score 3	Mean	Median
Original Dataset (66,005 entries)					
Linguistic Quality	1772 (2.7%)	48399 (73.3%)	13787 (20.9%)	2.19	2.0
Factual Accuracy	3071 (4.7%)	28003 (42.4%)	32884 (49.8%)	2.47	3.0
Instruction Adherence	288 (0.4%)	5435 (8.2%)	58235 (88.2%)	2.91	3.0
Helpfulness Relevance	80 (0.1%)	13501 (20.5%)	50377 (76.3%)	2.79	3.0
Filtered Dataset (59,242 entries)					
Linguistic Quality	N/A	45844 (77.4%)	13398 (22.6%)	2.23	2.0
Factual Accuracy	N/A	26799 (45.2%)	32443 (54.8%)	2.55	3.0
Instruction Adherence	N/A	4177 (7.1%)	55065 (92.9%)	2.93	3.0
Helpfulness Relevance	N/A	9368 (15.8%)	49874 (84.2%)	2.84	3.0
Human Evaluation Subset (100 entries)					
Linguistic Quality	07 (7.0%)	58 (58.0%)	35 (35.0%)	2.28	2.00
Factual Accuracy	01 (1.0%)	07 (7.0%)	92 (92.0%)	2.91	3.00
Instruction Adherence	02 (2.0%)	16 (16.0%)	82 (82.0%)	2.80	3.00
Helpfulness Relevance	00 (0.0%)	09 (9.0%)	91 (91.0%)	2.91	3.00

Table 3: LuxIT Score Distribution. We report scores from post-filtering on LuxIT with scores 1 (low), 2 (acceptable) and 3 (excellent). We show the score distribution on the original dataset compared with the filtered dataset after rejecting 6763 samples with low scores. We also report the score distribution for our manually evaluated subset of LuxIT.

trol, employing both an LLM-as-a-judge approach and manual human evaluation, confirmed that the majority of the generated pairs were of high quality. This result is significant as it presents a viable blueprint for creating instruction tuning resources in other low-resource languages, moving beyond the common practice of translating existing English datasets. By sourcing from native Luxembourgish texts from Wikipedia and RTL news, we aimed to reduce hallucination during data generation and create a culturally and linguistically authentic resource.

Our fine-tuning experiments conducted for RQ2 show mixed results. Out of the 3 smaller models we evaluated, Llama-3.2-1b and Gemma-3-1b scored notably worse compared to Qwen-3-0.6b which is surprising, as we expected larger models to perform better on average. Rather than simply answering questions wrong, both Gemma-3 and Llama-3.2 produced invalid outputs to the language exam questions.

As expected, the model that improved most across all levels overall, GLM-4-9b-0414, is also the largest one. Interestingly, Qwen-3-0.6b showed improvement on more levels than Llama-3.1-8b.

A compelling observation from our quality assessment is the subtle difference in rejection reasons between the LLM-as-a-judge and human evaluator. While GPT-5-mini was expected to be a primary filter for factual inaccuracies, our human

evaluator was more likely to identify issues related to linguistic nuance. This suggests that even advanced LLMs may not fully capture the idiomatic and grammatical subtleties of a low-resource language like Luxembourgish, highlighting the continued importance of human expertise in the evaluation pipeline.

## 7. Conclusion

This work addresses the scarcity of Luxembourgish resources for instruction fine-tuning large language models. Our research makes two primary contributions. First, we demonstrate that a Luxembourgish-proficient LLM can effectively generate a substantial, high-quality instruction-answer dataset from exclusively monolingual seed data, retaining 89.8% of the generated samples. Second, we investigate the dataset’s practical value by fine-tuning several smaller LLMs to measure changes in their Luxembourgish capabilities. The results of this benchmarking are mixed. While some models showed improvement on certain levels, others performed significantly worse. These findings suggest that further investigation, potentially with an expanded dataset, is necessary to better understand its impact and train more consistently performing models.

Our methodology provides a blueprint for creating culturally and linguistically authentic resources

in other low-resource settings, moving beyond a reliance on translating English-centric data. LuxIT serves as an important addition to the expanding repository of Luxembourgish language resources, providing a critical foundation for the future development of more proficient and accessible models for its speakers.

Building on this foundation, future work should focus on addressing the dataset’s current limitations to unlock its full potential. Given our mixed fine-tuning results, the most critical next step is to expand LuxIT at scale. This expansion should also aim to diversify the seed data beyond news and encyclopedic articles, incorporating other text forms like literature or parliamentary debates to capture a wider range of language styles. Furthermore, future efforts could move beyond single-turn, four-type instructions to include multi-turn conversational data and a greater variety of instruction types, which would be essential for training more capable and interactive Luxembourgish chatbots.

## 8. Limitations

We acknowledge several limitations. First, LuxIT, with its 59 242 entries is still relatively small compared to other instruction tuning datasets. Given our mixed results on the language exam evaluation, this scale may be insufficient for consistently training robust models.

Second, 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.

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

Fourth, our LLM-as-a-judge approach relied on a single model, GPT-5-mini, for quality scoring. While this model was chosen for its cost-performance trade-off, a single judge may not capture all errors, a concern supported by our human evaluation which identified more linguistic nuance. A more robust approach, such as using a majority-voting ensemble of several different judge models, might yield a higher-quality dataset, however, we retained the single-model approach as a necessary trade-off against the significant computational expense of an ensemble.

Fifth, 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.

gish.

## 9. Bibliographical References

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. [PromptSource: An integrated development environment and repository for natural language prompts](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva,INDERJIT Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, and Nan-Jiang Jiang et al. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#).

DeepSeek-AI. 2025. Deepseek-r1-0528 release. <https://api-docs.deepseek.com/news/news250528>. Accessed on August 18th, 2025.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyi Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao,



- and Aixin Liu et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#).
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#).
- 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](#).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#).
- Maxime Labonne. 2024. Fine-tune llama 3.1 ultra-efficiently with unsloth. <https://huggingface.co/blog/mlabonne/sft-llama3>. Accessed on August 18th, 2025.
- 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](#).
- Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. [Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation](#).
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. [The flan collection: Designing data and methods for effective instruction tuning](#).
- Cedric Lothritz and Jordi Cabot. 2025. [Testing low-resource language support in llms using language proficiency exams: the case of luxembourgish](#).
- Cedric Lothritz, Bertrand Lebigot, 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.
- 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 multi-task finetuning](#).
- Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. [Learning to generate instruction tuning datasets for zero-shot task adaptation](#).
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and Alex Paino et al. 2024a. [Gpt-4o system card](#).
- OpenAI. 2025a. [Gpt-5 system card](https://openai.com/index/gpt-5-system-card/). <https://openai.com/index/gpt-5-system-card/>. Accessed on August 18th, 2025.
- OpenAI. 2025b. [Openai o3 and o4-mini system card](https://openai.com/index/o3-o4-mini-system-card/). <https://openai.com/index/o3-o4-mini-system-card/>. Accessed on August 18th, 2025.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and Irwan Bello et al. 2024b. [Gpt-4 technical report](#).
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. [Instruction tuning with gpt-4](#).
- Fred Philippy, Laura Bernardy, Siwen Guo, Jacques Klein, and Tegawendé F. Bissyandé. 2025a. [Luxinstruct: A cross-lingual instruction tuning dataset for luxembourgish](#).

- 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](#).
- 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, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. [Multitask prompted training enables zero-shot task generalization](#).
- Manli Shu, Jiong Xiao 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.
- Qingyi Si, Tong Wang, Zheng Lin, Xu Zhang, Yanan Cao, and Weiping Wang. 2023. [An empirical study of instruction-tuning large language models in chinese](#).
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. <https://crfm.stanford.edu/2023/03/13/alpaca.html>. Accessed on September 16, 2024.
- 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](#).
- 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](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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](#).
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#).
- 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?](#)
- 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](#).
- 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.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chu-jie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and Jian Yang et al. 2025. [Qwen3 technical report](#).
- 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](#).

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu 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](#).

## 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<sup>11</sup> 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
```

<sup>11</sup>We do not show title, header and text for the RTL news article as the data is not publicly available

```
ginn ass..."
}
```

### A.2. Data generation model prompt

The data generation is performed by DeepSeek-R1-5028. 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",
    "response": "Detailed response in Luxembourgish"
  }
]
```

SOURCE TEXT:

```
{source_context}
```

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

### A.3. Prompt Template

Here we show the prompt template in more detail (Table 1). The Instruction and Response columns are brought together in the ShareGPT format:

#### Prompt Template

```
conversations =
[
  {
    "from": "human",
    "value": <instruction>
  },
  {
    "from": "gpt",
    "value": <response>
  }
]
```

```
]
```

### A.4. 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>  
}}
```