Text Generation Models for Luxembourgish with Limited Data: A Balanced Multilingual Strategy

Alistair Plum◇, Tharindu Ranasinghe♠, Christoph Purschke◇

◇University of Luxembourg, Esch-sur-Alzette, Luxembourg
♠Lancaster University, Lancaster, UK
{alistair.plum,christoph.purschke}@uni.lu
t.ranasinghe@lancaster.ac.uk

Abstract

This paper addresses the challenges in developing language models for less-represented languages, with a focus on Luxembourgish. Despite its active development, Luxembourgish faces a digital data scarcity, exacerbated by Luxembourg's multilingual context. We propose a novel text generation model based on the T5 architecture, combining limited Luxembourgish data with equal amounts, in terms of size and type, of German and French data. We hypothesise that a model trained on Luxembourgish, German, and French will improve the model's cross-lingual transfer learning capabilities and outperform monolingual and large multilingual models. To verify this, the study at hand explores whether multilingual or monolingual training is more beneficial for Luxembourgish language generation. For the evaluation, we introduce *LuxGen*, a text generation benchmark that is the first of its kind for Luxembourgish.

1 Introduction

Recent advances in deep learning have made it extremely popular to use language models (LMs) such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) for many tasks in natural language processing (NLP) (Lin et al., 2022). The tasks range from text classification tasks, such as sentiment analysis (Zhang et al., 2024b) and offensive language detection (Zampieri et al., 2023), to text generation tasks, such as machine translation (Zhu et al., 2024; Wang et al., 2023) and text summarisation (Zhang et al., 2024a; Liu et al., 2024). LMs have achieved state-of-the-art results in many of these tasks (Islam et al., 2024).

Even though these LMs are multilingual by design, their support and performance can suffer with languages that are not as well represented (Blasi et al., 2022). A lot of focus of these LMs tends to fall on English, as well as other high-resource languages (Pires et al., 2019). This fact is evidenced by the amount of data used for training certain models. For example, MT5 (Xue et al., 2021), which is trained using the Oscar common crawl data, contains roughly 3.4TB for English, 1.4TB for Chinese, 1.1TB for Russian, 600GB for German and 430GB for Spanish[1]. By contrast, Luxembourgish only has 18MB, and is therefore not well-supported by many current LLMs. However, initial testing has shown that GPT-4o (Achiam et al., 2023) models can produce Luxembourgish very well, demonstrating the positive effects of cross-lingual transfer-learning (Chen and Ritter, 2021) at an immense scale.

More data would certainly improve the situation, but this may not always be possible, meaning that other options need to be considered. Luxembourgish, a West Germanic language, is spoken by around 400,000 people, primarily in Luxembourg (Gilles, 2019). In addition to the small number of speakers, Luxembourg is home to a complex societal multilingualism that, historically, has favoured French and German as official languages, especially in formal and written communication. Only since the advent of digital and social media, and as a result of the active language policy to support Luxembourgish, have more significant amounts of text data been produced and therefore made available. As this situation is surely similar with other small varieties, means of finding and using more data are necessary.

Continuing with the example of Luxembourgish, models such as LUXEMBERT (Lothritz et al., 2022) have used data augmentation, in this case translating from German (Olariu et al., 2023). Models such as LUXGPT2 (Bernardy, 2022) rely on transfer learning. However, none of these methods have provided state-of-the-art performance like models for other languages. In fact, Ranasinghe et al. (2023) has shown that multilingual models

---

[1] https://oscar-project.github.io/documentation/versions/oscar-2301/

outperform existing monolingual Luxembourgish models in a text classification task. We, therefore, propose to fill this gap by combining the largest collection of data for the Luxembourgish language available with carefully considered transfer learning. In this case, in particular, we seek to answer the question of whether similar languages included in the training can improve performance. Recent research has shown that careful selection of multilingual training data improves models for 16 African languages (Oladipo et al., 2023). We therefore hypothesise that equal amounts, in terms of size and composition, of Luxembourgish, German, and French will outperform monolingual and multilingual models on Luxembourgish NLP tasks.

In this paper, we address the gap in Luxembourgish NLP and present a novel generative model that is based on the multilingual T5 architecture (Xue et al., 2021). Moreover, we run multiple evaluations on downstream tasks to ascertain whether multilingual or monolingual pre-training is more beneficial for a Luxembourgish model. To this end, we obtain training data for both German and French, the geographical and socio-cultural neighbours of Luxembourg(ish), and aim to learn more about treating these languages equally in pre-training. In doing so, we also discuss the data compilation, specifically the equivalency of data. For evaluation purposes, due to the current under-represented situation of Luxembourgish NLP, we re-use classification tasks (Lothritz et al., 2022) and introduce new generative tasks, including news article headline generation, paraphrasing and Wikipedia biography summaries.

Our **main contributions** are:

1. Two new generative language models for Luxembourgish, one pre-trained with just Luxembourgish, one pre-trained with Luxembourgish, German and French textual data.[2]

2. The introduction of *LuxGen*, the first text generation benchmark for Luxembourgish, including four new text generation tasks.

3. An evaluation of various language models in terms of performance in Luxembourgish.

4. Valuable insights into training data composition to increase the performance of a low-resource language.

---

[2]All material will be available via https://huggingface.co/instilux

2 Related Work

The expansion of language models to encompass European languages beyond English has been a focal point in recent NLP research. The T5 (Raffel et al., 2020), BERT (Devlin et al., 2019), and ELECTRA (Clark et al., 2020) architectures have been adapted to create not just multilingual (Xue et al., 2021) but also monolingual models that capture the linguistic nuances of individual languages. For example, Chan et al. (2020) developed GBERT and GELECTRA, and later GERMANT5, a version of T5 pre-trained exclusively on a large German corpus. Similarly, Le et al. (2020) introduced FLAUBERT, a French language model based on the BERT architecture.

In addition to these, Carmo et al. (2020) introduced PTT5, a Portuguese T5 model that outperformed previous models in Portuguese text generation and comprehension tasks. Araujo et al. (2024) trained a similar T5 model for Spanish, and Daðason and Loftsson (2022) evaluated various classification tasks on monolingual Icelandic models based on ELECTRA and CONVBERT (Jiang et al., 2020). For the Russian language, Zmitrovich et al. (2024) developed RUT5, among many others, achieving new benchmarks in Russian NLP tasks. These models underline the importance of tailoring language models to specific languages to capture their unique syntactic and semantic properties.

As a relatively small language, Luxembourgish is less represented in NLP, particularly in comparison to French and German, its socio-cultural neighbours. Luxembourgish has developed in close contact with French and German and today shares grammatical features as well as parts of the lexicon with those languages. This is especially true of German, due to Luxembourgish having developed from the Moselle Franconian dialect (Gilles, 2019), but is also true of French. While not typologically, Luxembourgish has been in close contact with French, exhibiting the borrowing of grammatical structures and lexical items, as well as a lot code-switching in written texts. Research in NLP for Luxembourgish has started only recently, with some early exceptions: Adda-Decker et al. (2008) introduce various resources for NLP tasks for Luxembourgish; Snoeren et al. (2010) analyse typical writing patterns (contextual n-deletion) in written transcripts of speech, and Lavergne et al. (2014) introduce a manually annotated corpus of mixed language sentences to test a word-based language

identification system. Recently, Sirajzade et al. (2020) and Gierschek (2022) introduced a state-of-the-art pipeline for sentiment analysis, Purschke (2020) published a pipeline for the automatic orthographic correction of text data and Philippy et al. (2024) introduced a new approach to Zero Shot Classification based on a task-specific dictionary for topic classification. For spoken data, Gilles et al. (2023a) and Gilles et al. (2023b) develop LUX-ASR, a performant Automatic Speech Recognition for Luxembourgish. For automatic comment moderation (Ranasinghe et al., 2023) and orthographic normalization (Lutgen et al., 2024), models based on the T5 architecture have shown to perform well for Luxembourgish.

Looking at language models for Luxembourgish, various strategies have been tested. Some have been trained using transfer-learning from German, such as LUXGPT (Bernardy, 2022). Another model for Luxembourgish exists in LUXEMBERT (Lothritz et al., 2022), which used augmentation techniques to produce more Luxembourgish data, and which performs on par with multilingual BERT on Luxembourgish language tasks such as part-of-speech tagging, named entity recognition and news classification (Lothritz et al., 2023). To this end, the authors of the afore mentioned model also introduced the corresponding resources. Anastasiou (2022) introduced ENRICH4ALL, another BERT model for the development of a multilingual chatbot in administrative contexts, trained with a specifically annotated corpus. These models all demonstrate the various strategies used to create the models.

Nevertheless, it is clear that not enough resources for Luxembourgish exist yet, and that research related to NLP in Luxembourgish is limited. Both of these reasons explain why there are only some benchmarks for classification tasks in Luxembourgish, and none for generative tasks. The absence of benchmarks, small amounts of data, and lack of more generative models are areas we hope to address with this research.

3 Data

Data is essential to training a language model and plays a critical role in this research, as outlined in the introduction. Because we wanted to investigate specifically the composition of the data used for training, we explain the choices made in this section.

It would be reasonable to assume a typical approach of simply using all the data one can find. Because Luxembourgish only has a minimal amount of data available compared to German and French, this would lead to a large imbalance in the data, calling into question whether the Luxembourgish part would even be worth including. As stated previously, we therefore aimed to match the German and French data largely to mirror that of the Luxembourgish data in terms of size, type, and domain. We consciously set out to collect roughly equal amounts of data for each language to test specifically how much better the language model would perform in comparison to a monolingual Luxembourgish model. Table 1 presents an overview of the different areas and the number of tokens from these areas, which will be described in the following paragraphs for each language.

Domain	LB	DE	FR
Radio	17,5M	1,3M	-
News	42,5M	71,7M	62,1M
Parl.	17,4M	31,2M	40,0M
Web	84,9M	56,2M	83,0M
Wiki	7,4M	16,4M	18,6M
Total	**169,7M**	**176,8M**	**203,7M**

Table 1: Token counts for texts from each domain for each language covered in the training data selection.

Luxembourgish For Luxembourgish, we opted to compile the dataset ourselves, as opposed to just using crawl data. This is due to reasons pointed out in previous sections, as well as affording us more control over the incoming data and allowing a more controlled compilation of corresponding data in the other languages. To this end, we aimed to collect all data as it is available to us (un-normalized) from known collections. This includes news articles (News), transcribed radio interviews (Radio) and user comments (Coms.) from the country's largest public news broadcaster, RTL Lëtzebuerg (RTL). As seen in Table 2, this forms the bulk of the Luxembourgish data in terms of the number of tokens. The RTL data spans the years 2008 until 2023 and has been used for many other Luxembourgish NLP research (Lothritz et al., 2022, 2023; Gierschek, 2022; Purschke, 2020; Ranasinghe et al., 2023). In terms of news related text, we also added data from the Leipzig (Lpzg) collection (Goldhahn et al., 2012) which includes data from other Luxembour-

gish news sites.[3]

For web content, we also made use of data from the Leipzig collection using specifically 1 million sentences from the latest web crawl, excluding RTL. In addition, we use text from Luxembourgish chat rooms. Encyclopaedic text was also used in the form of Wikipedia articles, which we obtained from the latest dump as of the time of training. This spans roughly 70,000 articles about various topics, but mainly biographies and information about locations. Similarly, we used all example sentences from the Luxembourgish online dictionary (LOD).[4] Finally, political speeches and debates are also represented in the corpus, which are transcribed for the Chambre des Députés (Chamber), the national legislature chamber of Luxembourg.

Resource	Tokens	Types	TTR
RTL Radio	17,5M	741,000	.0423
RTL News	36,7M	1,46M	.0398
RTL Coms.	55,8M	2,58M	.0463
Lpzg. News	5,85M	677,000	.1158
Lpzg. Web	17,1M	1,83M	.1074
Chat Logs	12,1M	659,000	.0545
Chamber	17,4M	404,000	.0233
Wikipedia	6,87M	576,000	.0839
LOD	0,5M	44,000	.0874
Total	**169,73M**	**8,87M**	**.0628**

Table 2: Token counts, type counts, and Type-Token Ratio (TTR) for Luxembourgish language resources.

German For German, we aimed to collect the closest corpora to the Luxembourgish corpora. Starting with news related text, we again made use of parts of the Leipzig corpus, specifically from German news sites (in roughly equal token quantities) as well as the Ten Thousand German News Articles Dataset (10kGNAD) dataset[5], which consists of over 10,000 articles from an Austrian newspaper across nine topics. As such, these articles are the unused part of the One Million Posts Corpus (OMPC) (Schabus et al., 2017), which we use to replicate the situation in Luxembourgish, where we have news articles and the user discussion under each article. We also use the Potsdam Commentary Corpus (PCC) (Bourgonje and Stede, 2020) to add further user comments from German newspaper

websites. For the transcribed radio interviews in Luxembourgish, we found the closest equivalent in the German Radio Interviews (GRAIN) corpus (Schweitzer et al., 2018), which comprises a small amount of transcribed radio interviews.

For web content, we used the Leipzig Corpus web crawl data for German. We could not replicate chat room data for German, and decided to leave this, as it only makes up a small amount of the Luxembourgish data. For encyclopaedic text we naturally used Wikipedia again, selecting the Leipzig Corpus Wikipedia selection for ease of use, and since one year is almost equivalent to the whole Luxembourgish Wikipedia corpus. To add political speeches and debates, we used the German section of the Digital Corpus of the European Parliament (DCEP) (Hajlaoui et al., 2014), specifically the AGENDA, IM-PRESS, MOTION and REPORT subsections, as these contained the most relevant textual data.

French For French, we also aimed to collect the most related textual data; however, we found the situation to be somewhat different to German, with not as many resources easily available. We used again Leipzig Corpora for news, French News 2010 and 2022 1M sentences, as well as French Newscrawl 2020 1M sentences. In addition, we used French Mixed Typical 2012 1M sentences to represent typical web data. To supplement web data and more comment style content, we used the French Reddit dataset from Kaggle.[6]

As for both previous languages we used an extract of Wikipedia articles for encyclopaedic text, making use of the Leipzig Corpora yet again, the French Wikipedia 2021 1M sentences collection. For political speeches and discussion, we used the French section of the DCEP, with the same subsections as for German.

4 Models

We leverage our unlabelled data described in Section 3 to pretrain two models: LUXT5 on Luxembourgish data and LUXT5-GRANDE on Luxembourgish, German and French data using the T5-BASE encoder-decoder architecture (Raffel et al., 2020). Each of the encoder and decoder components contains 12 layers, each with 12 attention heads and 768 hidden units. In total, this results in a model with 220 million parameters.

---

[3]https://corpora.uni-leipzig.de/
[4]https://lod.lu
[5]https://tblock.github.io/10kGNAD/

[6]https://www.kaggle.com/datasets/breandan/french-reddit-discussion

Task	Train	Test	Type
News Title	162,882	13,852	RTL news articles
Positive comment	3,236	810	RTL articles & comments
Negative comment	3,236	810	RTL articles & comments
Description	11,858	2,094	Wikipedia articles

Table 3: Overview of the data for the four different LuxGen tasks, including no. of training and test instances, as well as the types of instances.

We used the same objective as the original T5 models (Raffel et al., 2020). The main idea is to feed the model with corrupted (masked) versions of the original sentence and train it to reconstruct the original sequence. This denoising objective works by randomly sampling and dropping out 15% of tokens in the input sequence. All consecutive spans of dropped-out tokens are then replaced by a single sentinel token.

For both of our pre-trained models, we use a learning rate of 1e-4, a batch size of 128 sequences, and a maximum sequence length of 512. We pre-train each model for 1M steps.

5 *LuxGen*: Text Generation Benchmark for Luxembourgish

To evaluate the LUXT5 and LUXT5-GRANDE models, we defined *LuxGen*: A text generation benchmark for Luxembourgish, consisting of four text generative tasks. Due to data being limited for Luxembourgish, especially in terms of benchmarks, we derive these tasks mainly from RTL data, as this already has the most metadata available. The generative tasks are all novel for Luxembourgish. We believe these tasks to be the best currently available to evaluate the performance of text generation models, including recent large language models for NLP tasks in Luxembourgish. An overview of the available data is presented in Table 3.

5.1 News Headline Generation

In this task, the model is trained to generate a headline for a specific news article. We created this task as it offered itself as a straight-forward task from the data. Similar tasks have been proposed by Hettiarachchi et al. (2024), Nagoudi et al. (2022), and Aralikatte et al. (2023).

We used news articles taken from the RTL collection that we used for pre-training the models. It should be made clear at this point that we removed the article headlines at the point of pre-training so that we could obtain unbiased results. The exact

number of training and testing instances can be seen in Table 3.

5.2 Positive and Negative Comment Generation

For this task, we utilise user voting on the RTL user comments dataset to extract the most upvoted and downvoted comments. Using the corresponding RTL article that a given user comment was made on, the task is to generate the most upvoted and the most downvoted comment.

The datasets used for this task are the RTL user comments dataset and the RTL news articles dataset. Matching the comments with the corresponding article by ID, we then calculate the up/down ratio and determine the most upvoted and most downvoted comments. Since the voting on user comments feature was only introduced in 2019, our data is partially limited for this task, especially as not every article has user comments that have votes. The comments have also been moderated by RTL to remove harmful or offensive language and anonymise users. We have 4044 comments each for most upvoted and most downvoted, utilising an 85% to 15% train test split in order to retain the maximum amount for training (see also Table 3).

5.3 Short Description Generation

We define the final generative task as description generation. We utilise Wikipedia and its structured equivalent, Wikidata, for this task. The task for the model is to generate a short description of a Wikipedia article.

For this task, we use all Luxembourgish Wikipedia articles that have a short description on Wikidata. These descriptions can almost be seen as short labels; nevertheless, we use this data for a generative task. As the number of articles in Luxembourgish is quite small, we collected roughly 14,000 articles with descriptions. An exact overview of the training and test instances is presented in Table 3.

Group	Model	Headline	Positive	Negative	Wiki
Prompt	GPT-4O-2024-05-13	0.0482	0.0032	0.0017	0.1001
	LLAMA-3.1-8B-INS.	0.0359	0.0037	0.0028	0.0268
Pre + Fine	LUXT5-GRANDE	**0.2130**	**0.0810**	**0.0780**	**0.1100**
	LUXT5	0.1680	0.0450	0.0320	0.0280
Fine-tuning	MT5-BASE	0.1820	0.0009	0.0006	0.0230
	MT5-SMALL	0.1650	0.0003	0.0003	0.0160
	BYT5-BASE	0.0310	0.0000	0.0000	0.0002
	BYT5-SMALL	0.0320	0.0000	0.0000	0.0001

Table 4: BLEU scores for different tasks in *LuxGen*. The best result for each task is in bold.

6 Evaluation

In this section, we evaluate the performance of our proposed models, LUXT5 and LUXT5-GRANDE, on Luxembourgish text generation tasks encompassed in the *LuxGen* dataset, and a classification task for Luxembourgish introduced by Ranasinghe et al. (2023). We compare our models against several baselines, including the non fine-tuned large language models (LLMs) LLAMA 3 (Dubey et al., 2024), GPT-4O (Achiam et al., 2023), and MISTRAL, as well as fine-tuned versions of mT5 (Xue et al., 2021) and BYT5 (Xue et al., 2022). Our evaluation comprises both automatic metrics, using BLEU scores (Papineni et al., 2002) due to the lack of advanced NLG metrics for Luxembourgish, standard metrics for classification, and a manual analysis to provide a comprehensive understanding of each model's capabilities in generating accurate and fluent Luxembourgish text.

6.1 LuxGen

For all tasks in *LuxGen*, we compare LUXT5 and LUXT5-GRANDE to Llama 3 and GPT 4o (non fine-tuned), as well as several fine-tuned variants of mT5 (Xue et al., 2021) and ByT5 (Xue et al., 2022). All the tasks were considered sequence-to-sequence tasks. For all the T5-based models, we used the same configurations: a batch size of 8, Adam optimiser with learning rate 1e-4, and a linear learning rate warm-up over 10% of the training data and trained the models over ten epochs. For Llama 3 and GPT 4o, we used prompts in English and optimised them to achieve the best output. Exact prompts are listed in Appendix A. MISTRAL did not produce any outputs in Luxembourgish.

For the automatic evaluation, we utilised BLEU score (Sharma et al., 2019). While there are advanced NLG metrics such as BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2019),

they do not currently support Luxembourgish. The results are shown in Table 4. As can be seen in Table 4, LUXT5-GRANDE outperforms LuxT5 and other baselines in all the tasks in *LuxGen*. The key findings of the results are listed below.

LUXT5-GRANDE As stated previously, LUXT5-GRANDE outperforms all models in the *LuxGen* tasks. For the tasks with more training instances, such as headline generation, the gap between the mT5 models and LUXT5-GRANDE is low. However, for the tasks where the number of training instances is lower, there is a larger gap between LUXT5-GRANDE and the mT5 models. We believe that when there are a large number of training instances, MT5 can come close to specific T5 models. However, they are unable to train their weights properly when there are fewer training instances. We also see that the LLMs do not perform as well, except GPT-4O in the Wikipedia description task, which could very well be due to the overlap of training data, i.e. GPT having seen Wikipedia in training.

Monolingual Training LUXT5, which we only trained using Luxembourgish data, does not consistently outperform mT5 models in *LuxGen*. We believe that this shows the Luxembourgish data on its own is simply insufficient. Since there was not much data to train LUXT5, the model might be inconsistent in some tasks. This is also shown by the LLM results, which do not reach the performance of LUXT5, but come close even without fine-tuning (but having many times more data in pre-training).

BYT5 models Previous research suggested that ByT5 models will perform well in Luxembourgish tasks (Ranasinghe et al., 2023). Surprisingly, the results in Table 4 suggest that ByT5 models perform poorly in Luxembourgish text generation tasks. We

Model	Archived			Published			Weighted Average			F1 Macro
	P	R	F1	P	R	F1	P	R	F1	
MBERT	0.58	0.06	0.12	0.77	0.97	0.86	0.72	0.77	0.70	0.49
LUXEMBERT	0.60	0.08	0.15	0.78	**0.98**	0.87	0.73	0.77	0.70	0.51
BYT5 LARGE	0.67	**0.20**	**0.31**	**0.79**	**0.98**	**0.88**	**0.77**	0.78	**0.74**	**0.59**
LUXT5-GRANDE	**0.69**	0.15	0.25	**0.79**	**0.98**	**0.88**	**0.77**	**0.79**	0.73	0.56

Table 5: Results for the moderation classification task.

assume that this is due to the model architecture not being as well suited to generative tasks in the *LuxGen* settings.

6.2 Classification Evaluation

The classification task we evaluate involves predicting whether a given user comment is *Archived* or *Published*, as first introduced by Ranasinghe et al. (2023), for which we reproduced the same data splits for comparability. This task presents significant challenges due to class imbalance and the subtle distinctions between the two categories, making it a valuable benchmark for assessing model performance on Luxembourgish text classification.

The results presented in Table 5 indicate that our proposed model, LUXT5-GRANDE, outperforms the previously released Luxembourgish modelLUXEMBERT (Lothritz et al., 2022) across multiple evaluation metrics. Specifically, LUXT5-GRANDE achieves higher precision for the "Archived" class and a better weighted average precision. While BYT5 LARGE attains the highest overall performance with an F1 Macro score of 0.59, LUXT5-GRANDE comes close with an F1 Macro of 0.56, despite not being optimized for classification tasks.

We attribute the superior performance of BYT5 LARGE in this classification task to its architecture, which is well-suited for handling character-level information — a critical factor for languages with rich morphological structures like Luxembourgish. The byte-level tokenization employed by BYT5 LARGE enables it to capture subtle textual nuances essential for accurate classification. In contrast, LUXT5-GRANDE was primarily developed as a generative model. Despite this, its competitive performance underscores the effectiveness of our pre-training and fine-tuning approach. With further optimization targeted at classification, LUXT5-GRANDE has the potential to surpass existing models by combining strong generative capabilities with robust classification performance.

6.3 Manual Evaluation

Due to the fact that the BLEU score does not offer complete insight into the performance of the models for Luxembourgish, we also completed a manual investigation and evaluation of some sample sentences to gain more insight into the generated text. As we observed during the evaluation, the two evaluated LLMs produced seemingly good Luxembourgish, although often having sentences with reversed logic to that of the real test output (for example, *team B lost to team A* versus *team A won against team B*). Because of BLEU's evaluation on word alignment, this impacts the scores heavily. Therefore, we have opted to include our analysis of a sample of the output predictions for *LuxGen*. Since our pre-trained LUXT5-GRANDE model performed best in terms of BLEU out of the T5-based models, we selected its output, alongside the monolingual LUXT5 to directly compare the effects of adding more languages in pre-training, as well as the two LLMs, for further analysis.

For the analysis, we took 20 random sentences per task in *LuxGen*, and evaluated the predictions by taking three categories into account: *task*, *content*, and *correctness*. With *task*, we checked whether the task had been completed, e.g. *has a headline been generated? Is the length appropriate?*. For *content*, we compared not only the target text but also the input text to check whether the model has reproduced appropriate content or not. Finally, for *correctness*, we checked the output according to Luxembourgish grammar and orthography rules, as well as whether the model stayed in the correct language. We summarise our findings per model group.

LLMs As part of the evaluation, we prompted GPT-4O and LLAMA-3.1-8B-INSTRUCT to generate predictions for *LuxGen*. Looking at the outputs, we consistently saw that both models were able to complete the assigned tasks, which is to be expected. It was observed that although both models

tended to generate texts that were longer than the target outputs (see example outputs below), the tasks were still completed, usually displaying more information than the targets. This relates also to content, which was generally reproduced correctly, although headlines tended to contain much more information than the target headlines, but the articles did contain this information. In terms of correctness, GPT-4o was mainly able to produce correct Luxembourgish, only rarely switching to German or hallucinating Luxembourgish forms, as indicated in the example outputs in red. LLAMA-3.1-8B-INSTRUCT, on the other hand, did not produce correct Luxembourgish, often switching midway through predictions to German, highlighted in the example sentence below in blue. Compared with the other two models, the LLMs produced more passive constructions than active ones. On the whole, it is fair to say the BLEU scores do not accurately reflect the quality of the output of these models in terms of task and language.

Example Outputs:

- `Llama:` Bolivien: Dausende Polizisten jagen international gesuchten Drogeboss Sebastian Cabrera.

- `GPT:` Dausende Polizisten an Bolivien op déi grofl Botter - International gesichte Drogeboss entkommt nach ëmmer.

- `LuxT5:` Dausende Poliziste kämpfe géint Drogeboss Sebastian Cabrera.

- `Original:` Police sicht no Drogeboss Sebastian Cabrera.

LuxT5s Looking at the results of our LUXT5 models, we saw that both models were able to complete the various tasks. In fact, both models generated outputs that were much more similar in length and style to the target outputs, which is to be expected due to these models being fine-tuned for the various tasks. In terms of content, we often found that LUXT5 would often add random bits of information that it did not reproduce from the text inputs, making it factually incorrect in places. LUXT5-GRANDE did not suffer from this, demonstrating that adding more language data in addition to the Luxembourgish base to be beneficial. We also saw that both models do not switch around the sentence logic, as observed with the LLMs, but did slightly change the meaning, as highlighted in the example outputs above in gold. Finally, in terms

of correctness, both models generated Luxembourgish without switching to German, with only minor mistakes. It should be noted, however, that both models suffered slightly from finishing the generation too early, therefore leaving unfinished words in places.

Overall We saw that all four models produced much better output than the automatic evaluation would indicate. It seems clear that this has much to do with the fact that the outputs, while addressing the task, being factually correct, and linguistically correct, often look nothing like the target predictions. Because this would mean that many words are misaligned, the BLEU scores would suffer from this. Although this is not optimal, the targets that we have are the only ones that we can work with that have been produced by real humans. Nevertheless, with these results, we see that including similar languages in the pre-training process can improve the performance in language models.

7 Conclusion

This paper has demonstrated the performance of multiple language models for the Luxembourgish language. The models demonstrated are all capable of both text classification and text generation tasks. While we have presented a detailed evaluation of the various models, we have also described the different datasets with which the models were trained. In doing so, we have shown that large, massively multilingual models do not necessarily perform best for small, low-resource languages. In fact, we think it is clear that smaller models, trained on the limited amount of data available for a given low-resource language, can benefit from the addition of equal amounts of linguistically related languages. Our results indicate that such models can outperform larger multilingual language models consistently and can come very close to the performance of LLMs, like GPT, although at a considerably lower cost in terms of training data size and training time. With these findings in mind, we plan ablation studies in the future to determine the exact effects more precisely.

The findings of this paper further suggest that there may be positive implications for not just low-resource languages that can benefit from socio-cultural neighbours or contact languages, but also for all kinds of varieties within a given language, such as regional dialects. Nonetheless, these findings require further research, which will shape our

future outlook on the topic of this paper. We plan to experiment with adding and removing further languages to and from our models to assess the performance impact. We also want to look more closely at the precise quantities and composition of added data, as well as the balances in relation to other languages. Furthermore, we plan to test our approach for regional varieties of German and other languages, to determine whether this approach of adding linguistically related languages or language varieties to a model can help performance.

Limitations

It is clear that using BLEU score is not an ideal metric, especially given the fact that Luxembourgish is not widely standardised in practice, meaning that character variation is always present, making a character-based metric difficult to interpret for evaluation. However, due to Luxembourgish being a low-research language, we could not determine a more suitable metric. The fact that there is limited data for Luxembourgish, and that there is only a tiny amount of human annotated data, exacerbates this problem.

Ethics Statement

This research was conducted using existing annotated datasets and did not involve the creation of any new human annotations. All data utilized in this study was previously publicly available and did not require any new data collection. The datasets employed in this research are properly licensed for this use.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of "Lëtzebuergesch" Resources for Automatic Speech Processing and Linguistic Studies. In *Proceedings of LREC 08*, Marrakech, Morocco. ELRA.

Dimitra Anastasiou. 2022. ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot. In *Proceedings of ELRA/ISCA Special Interest Group on Under-Resourced Languages*, Marseille, France. ELRA.

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. Varta: A Large-Scale Headline-Generation Dataset for Indic Languages. In *Findings of ACL 2023*, Toronto, Canada. ACL.

Vladimir Araujo, Maria Mihaela Trusca, Rodrigo Tufiño, and Marie Francine Moens. 2024. Sequence-to-Sequence Spanish Pre-trained Language Models. In *Proceedings of LREC-COLING 2024*.

Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master's thesis, University of Trier.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of ACL*, Dublin, Ireland. ACL.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of LREC 2020*.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PtT5: Pre-training and validating the T5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of ICCL*.

Yang Chen and Alan Ritter. 2021. Model Selection for Cross-lingual Transfer. In *Proceedings of EMNLP 2021*, Punta Cana, Dominican Republic. ACL.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *CoRR*, abs/2003.10555.

Jón Friðrik Daðason and Hrafn Loftsson. 2022. Pre-training and Evaluating Transformer-based Language Models for Icelandic. In *Proceedings of LREC 2022*, Marseille, France. ELRA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019: HLT*, Minneapolis, Minnesota. ACL.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.

Peter Gilles. 2019. 39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*. De Gruyter Mouton, Berlin, Boston.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023a. ASRLUX: Automatic Speech Recognition for the Low-Resource Language Luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023b. LUX-ASR: Building an ASR system for the Luxembourgish language. In *Proceedings - 2022 IEEE Spoken Language Technology Workshop (SLT)*.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, Daniel Varga, et al. 2014. DCEP-Digital Corpus of the European Parliament. In *LREC*.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A News Corpus for Sinhala. In *Proceedings of LREC-COLING 2024*, Torino, Italia. ELRA and ICCL.

Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Nagat Drawel, Gaith Rjoub, and Witold Pedrycz. 2024. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Systems with Applications*, 241.

Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. *NeurIPS*, 33.

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish. In *Proceedings of LREC 2014*, Reykjavik, Iceland. ELRA.

Hang Le, Loıc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoıt Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of LREC 2020*.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On Learning to Summarize with Large Language Models as References. In *Proceedings of NAACL: HLT*, Mexico City, Mexico. ACL.

Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé François D Assise Bissyande, Jacques Klein, Isabella Olariu, Andrey Boytsov, Clement Lefebvre, and Anne Goujon. 2023. Comparing Pre-Training Schemes for Luxembourgish BERT Models. In *Proceedings of KONVENS 2023*.

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 LREC 2022*, Marseille, France. European Language Resources Association.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2024. Neural Text Normalization for Luxembourgish using Real-Life Variation Data. *Preprint*, arXiv:2412.09383.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. AraT5: Text-to-Text Transformers for Arabic Language Generation. In *Proceedings of ACL 2022*, Dublin, Ireland. ACL.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better Quality Pre-training Data and T5 Models for African Languages. In *Proceedings of EMNLP 2023*, Singapore. Association for Computational Linguistics.

Isabella Olariu, Cedric Lothritz, Tegawendé Bissyandé, and Jacques Klein. 2023. Evaluating Data Augmentation Techniques for the Training of Luxembourgish Language Models. In *Proceedings of KONVENS 2023*, Ingolstadt, Germany. Association for Computational Lingustics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, USA. ACL.

Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, Use a Dictionary: Zero-Shot Topic Classification for Low-Resource Languages with Application to Luxembourgish. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages LREC-COLING 2024*, Torino, Italia. ELRA and ICCL.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of ACL*, Florence, Italy. ACL.

Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140).

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or Hold? Automatic Comment Moderation in Luxembourgish News Articles. In *Proceedings of RANLP 2023*, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One Million Posts: A Data Set of German Online Discussions. In *Proceedings of SIGIR 2017*, Tokyo, Japan.

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. German radio interviews: The GRAIN release of the SFB732 Silver Standard Collection. In *Proceedings of LREC 2018*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of ACL 2020*, Online. ACL.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.*, 10(3).

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL 2020*, Marseille. LREC.

Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda. 2010. The Study of Writing Variants in an Under-resourced Language: Some Evidence from Mobile N-Deletion in Luxembourgish. In *Proceedings of LREC 2010)*, Valletta, Malta. ELRA.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. In *Proceedings of EMNLP 2023*, Singapore. ACL.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *TACL*, 10.

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 NAACL: HLT 2021*, Online. ACL.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023. OffensEval 2023: Offensive language identification in the age of Large Language Models. *Natural Language Engineering*, 29(6).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024a. Benchmarking Large Language Models for News Summarization. *TACL*, 12.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024b. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of NAACL 2024*, Mexico City, Mexico. ACL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of NAACL 2024*, Mexico City, Mexico. ACL.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, Tatiana Shavrina, Vladislav Mikhailov, and Alena Fenogenova. 2024. A Family of Pretrained Transformer Language Models for Russian. *Preprint*, arXiv:2309.10931.

A List of Prompts

A.1 GPT-4o

- You are an editorial assistant for a Luxembourgish news outlet. Your task is to generate a news headline for a news article, based on the content of the article.

- You are a Luxembourgish social media user. Your task is to generate a positive user comment in response to a news article. The comment should be closest to a comment that is most likely to get the most upvotes or thumbs up from other users.

- You are a Luxembourgish social media user. Your task is to generate a user comment in response to a news article. The comment should be closest to a comment that is most likely to get the most downvotes or thumbs down from other users.

- Based on a Luxembourgish Wikipedia article as input, your task is to generate a short description in Luxembourgish of the thing that

is being described. The description can be as short as a word, and no longer than a short sentence.

A.2 Llama 3

- You are an editorial assistant for a Luxembourgish news outlet. Your task is to generate a news headline for the following news article, based on the content of the article. Only return the title.

- You are a Luxembourgish social media user. Your task is to generate a user comment in response to a news article. The comment should be closest to a comment that is most likely to get the most upvotes or thumbs up from other users. Only return the comment.

- You are a Luxembourgish social media user. Your task is to generate a user comment in response to a news article. The comment should be closest to a comment that is most likely to get the most downvotes or thumbs down from other users. Only return the comment.

- Based on a Luxembourgish Wikipedia article as input, your task is to generate the corresponding short Wikipedia description in Luxembourgish of the thing that is being described. The general description should not be longer than a couple of words. Only return the description.