A Fine-Grained Linguistic Evaluation of Low-Resource Luxembourgish–English MT

Nils Rehlinger
University of Luxembourg / Esch-Belval, Esch-sur-Alzette, Luxembourg
nils.rehlinger@uni.lu

Abstract

Machine translation (MT) evaluation is central in guiding researchers on how to improve a model's performance. Current automatic evaluation practices fail to provide reliable insights into the specific translation errors that occur, especially for low-resource languages. This paper introduces the Lux-MT-Test-Suite, enabling a linguistically motivated and fine-grained analysis of Luxembourgish–English (LB-EN) MT based on 896 test items covering 12 linguistic categories and 36 linguistic phenomena. We compare a baseline local LLM (GEMMA 3), its fine-tuned counterpart (LUXMT), and a proprietary state-of-the-art LLM (GPT-5) to analyse what local LLMs learn through fine-tuning in a low-resource setting and to assess performance differences between local and proprietary systems. The findings identify specific performance gains through fine-tuning, minor degradations, a difference in translation strategies, performance gaps between local and proprietary models, and remaining challenges.

1 Introduction

Machine translation (MT) evaluation is a key step in developing MT models. Its purpose is to guide researchers in discerning which strategies improve a model's performance. The standard procedure for MT evaluation is to assess models on a benchmark consisting of largely random sentences, e.g., FLORES-200 (Costa-Jussà et al., 2022; Manakhimova et al., 2025). However, this procedure fails to provide any information about what kind of errors occur, limiting its utility for diagnosing model weaknesses. As a result, there is a growing interest in evaluation methods that help identify MT errors in a fine-grained manner (Kocmi et al., 2025).

Against this backdrop, this paper introduces the Lux-MT-Test-Suite[1], the first fine-grained test suite for Luxembourgish–English (LB-EN) MT. The test suite consists of a diverse set of linguistically motivated test items that target LB-specific linguistic phenomena. These phenomena are then grouped together into broader linguistic categories.

LB is a West-Germanic language spoken by approximately 320'000 speakers (Fehlen and Heinz, 2016; Entringer et al., 2021). It is one of three official languages in Luxembourg, alongside German (DE) and French (FR). The language is closely related to DE but is characterised by frequent borrowing from FR. Historically, LB was restricted to the spoken domain. Only recently has the language been increasingly used in the written domain, e.g., on news and government websites. Due to the small number of speakers and domain restriction, parallel corpora are rare and LB can be considered as a low-resource language in the MT field.

Using the Lux-MT-Test-Suite, this paper analyses what translation capabilities local LLMs acquire through fine-tuning in a low-resource setting and compare their performance with state-of-the-art (SOTA) proprietary LLMs. Thus, this paper addresses the following explorative research question: *What Luxembourgish linguistic knowledge do local LLMs acquire through fine-tuning?* To answer this question, we fine-tune a GEMMA 3 model on LB parallel corpora, compare its performance to its pre-fine-tuned baseline and a proprietary SOTA LLM (GPT-5), highlighting differences across linguistic categories in the Lux-MT-Test-Suite.

The contributions of this paper are threefold:

1. Introducing Lux-MT-Test-Suite: the first MT test suite for LB-EN.

2. Providing insights into which linguistic phenomena local LLMs learn through fine-tuning in a low-resource setting.

3. Identifying performance gaps between local

[1] https://github.com/greenirvavril/lux-mt-test-suite

Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026), pages 138–150
March 28, 2026

LLMs and SOTA proprietary LLMs by conducting a fine-grained evaluation.

2 Related Work

2.1 Automatic MT Evaluation Metrics

As mentioned above, the standard procedure is to evaluate MT systems on a set of largely random sentences. The procedure usually involves computing average scores using automatic evaluation metrics. These metrics are generally designed to correlate with human judgements on translation quality, e.g., BERTSCORE (Zhang et al., 2019), BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), etc. While these scores can be helpful to rank models, they are opaque with regards to specific translation errors influencing the score.

2.2 Automatic Span-level Error Annotation

Recent efforts have attempted to tackle this issue by developing metrics that automatically label translation errors, such as XCOMET (Guerreiro et al., 2024) and GEMSPANEVAL (Juraska et al., 2025). While this approach seems promising, it requires vast amounts of annotated data, which are not available for low-resource languages like LB. Moreover, the accuracy of current SOTA automatic error detection models still shows a significant gap with human performance (Lavie et al., 2025).

2.3 Multidimensional Quality Metric Framework

The Multidimensional Quality Metric (MQM) framework is a popular method designed to assess translation quality by labeling error types and assigning severity levels (Lommel et al., 2014, 2024). MQM consists of an elaborate taxonomy of error categories. However, these categories remain generalist, leaving language-specific grammatical phenomena unexplored.

2.4 MT Test Suites

Given these limitations, MT test suites are a promising complementary tool to diagnose MT systems' shortcomings in a fine-grained manner. Also known as challenge sets, MT test suites consist of a collection of curated test items targeting specific linguistic phenomena. MT test suites have existed since the early 1990s, although their popularity has fluctuated over time (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). The latest advances in MT saw substantial performance boosts, causing automatic metric scores to become saturated and leading researchers to call for more challenging and detailed evaluation methods (Proietti et al., 2025; Kocmi et al., 2025). As a result, MT test suites are regaining popularity, with recent examples including Macketanz et al.'s (2022a) DE-EN test suite, Avelino et al.'s (2022) Portuguese–English (PT-EN) test suite, and Manakhimova et al.'s (2025) Russian–English (RU-EN) test suite, to name a few.

2.5 Luxembourgish NLP

Over the past few years, LB has seen a growing presence in the NLP space. Contributions include instruction fine-tuning datasets (Philippy et al., 2025a; Valline et al., 2025), a treebank (Plum et al., 2024), BERT models fine-tuned for various tasks (Gierschek, 2022; Lothritz et al., 2022; Anastasiou, 2022), ASR models (Gilles et al., 2023b,a), a model for comment moderation (Ranasinghe et al., 2023), a normaliser (Lutgen et al., 2025), embeddings (Philippy et al., 2025b; Michail et al., 2025), and generative models, such as LUXT5 (Plum et al., 2025) and LUXGPT (Bernardy, 2022). Developments in MT include LETZ TRANSLATE (Song et al., 2023) based on OPUS-MT (Tiedemann and Thottingal, 2020), LETZ-MT based on GEMMA 2 3B (Song et al., 2025), and KI-IWWERSETZER[2], a model developed using the OPENNMT ecosystem (Klein et al., 2017).

So far, LB MT evaluation has been limited to reference-based metrics, which ignore the source and are therefore susceptible to quality issues in reference translations (Moghe et al., 2025). In addition, as mentioned earlier, these metrics fail to identify translation errors. The following test suite seeks to address this gap by evaluating MT systems across a plurality of linguistically motivated test items representing LB-specific linguistic categories and phenomena.

Since most previous works in MT used automatic evaluation metrics that leave translation errors unidentified, the question of what exactly local LLMs learn during fine-tuning remains to a large extent under-researched. To address this gap, this paper introduces the Lux-MT-Test-Suite and uses it to evaluate and compare a baseline local LLM (GEMMA 3) with a fine-tuned counterpart, as well as a SOTA proprietary LLM (GPT-5). Through

[2] https://iwwersetzung.lu

this evaluation and comparison, we explore what local LLMs learn through fine-tuning in a low-resource setting and identify performance gaps between local and proprietary LLMs.

3 Method

3.1 Model Selection

While local MT systems for LB exist (see Section 2.5), the models either do not support LB-EN translations or they are out-dated. For this reason, this paper focuses on a fine-tuned version of GEMMA 3 that is currently under development for an on-going MT project (Team et al., 2025). The base model is the instruction fine-tuned 8-bit quantised version with 27 billion parameters. While larger and possibly more performant local models exist, GEMMA 3 is among the largest models that we could run on our hardware. As for the proprietary LLM, we choose the popular GPT-5 model[3] and access it via its API.

3.2 Data Preparation & Fine-tuning

The GEMMA 3 model was fine-tuned using the Unsloth[4] fine-tuning suite. As data, we used `LuxAlign` (Philippy et al., 2025b), consisting of 89k LB-FR and 28k LB-EN segment pairs from RTL News[5]. The reason to also include FR data in the mix is to benefit from cross-lingual transfer learning (Philippy et al., 2023). Since RTL articles are not 1-to-1 translations, LUXEMBEDDER was used with a cosine similarity threshold of .99 to filter out low-equivalence segment pairs, reducing the parallel corpus to 14k segment pairs for LB-FR and 2.5k for LB-EN. We augmented the data using Google Translate[6] to translate LB parliamentary debate transcripts to EN and FR. The augmented data was also filtered using LUXEMBEDDER with a threshold of .98, supplementing the dataset with an additional 20k segment pairs for LB-EN and 18k for LB-FR. All data was checked for duplicates and the minimum segment-length was 5 words. The hyperparameters include a learning-rate of 2e-5 and the model was fine-tuned for one epoch.

3.3 Test Suite Creation

The Lux-MT-Test-Suite is largely based on Macketanz et al.'s (2022a) test suite for DE-EN due to the linguistic similarity between LB and DE. We selected a set of DE source sentences from the DE-EN test suite to translate into LB. The selection was guided by how suitable the source sentences would be for translation while maintaining the grammatical structures relevant to the targeted phenomenon. Other sources include example sentences from the Luxembourgish Online Dictionary[7] (LOD), sentences from Döhmer (2020) and some sentences were also devised by the present author who is a native LB speaker with a background in linguistics.

The examples are designed to require translations that involve restructuring in the target language, thereby posing a challenge to MT systems (Manakhimova et al., 2025).

3.4 Test Suite Overview

The test suite contains 896 test items covering 12 linguistic categories, which are subdivided into a total of 36 linguistic phenomena. Each phenomenon consists of at least five test items (see Table 1 for a category-level overview and Table 3 in Appendix A for a phenomenon-level overview).

The category ***Ambiguity*** contains test items in which a lexeme has multiple possible meanings, requiring the MT system to disambiguate the meaning from the context. ***Coordination & ellipsis*** includes test items containing different kinds of ellipsis that require MT systems to perform syntactical restructuring. ***False friends*** contains lexemes that have a form similar to a corresponding EN lexeme, but different meanings. ***Function word*** includes test items in which focus particles or question tags contribute to pragmatic nuances. ***LDD & interrogatives*** checks long-distance dependencies and related discourse phenomena. ***Lexical morphology*** covers noun formation through the nominalisation of verbs and adjectives, and gender variation in nouns. ***MWE*** (Multi-word entities) contains idioms, prepositional and verbal MWE, and collocations, i.e., (semi-)fixed combinations of lexical items. ***Named entity & terminology*** contains LB place names and festivities, and dates. ***Non-verbal agreement*** checks case and gender agreement between subjects and objects. ***Subordination*** assesses a range of subordinate clause constructions. ***Verb tense/aspect/mood*** checks if MT sys-

[3] `https://chatgpt.com/`, data collected on November 24th, 2025.

[4] `https://unsloth.ai/`

[5] The largest news media outlet in Luxembourg: `https://www.rtl.lu`

[6] Data augmented in August 2025.

[7] `https://lod.lu`

tems correctly handle verbal tenses and persons, including their correct auxiliary and model verbs, as well as verbal inflections. Lastly, ***Verb valency*** examines if MT systems correctly translates verbs with their fixed number of arguments.

Together, these categories target syntactical, morphological, pragmatic, and discourse-level features, providing a fine-grained analysis of difficulties in LB-EN MT.

3.5 Scoring

All test items include a set of evaluation rules in the form of (in-)correct tokens or regular expressions to automatically flag the candidate sentences as *correct* or *incorrect*. The tokens are fixed strings of translated test items representing correct or incorrect solutions. The evaluation rules are either transferred from (Macketanz et al., 2022a) or manually written.

The evaluation process is semi-automatic: MT outputs are flagged for correctness based on the pre-written evaluation rules. Items that fail to be flagged are manually evaluated. This is usually the case for novel MT outputs that were not previously captured by the evaluation rules. The new manual evaluation annotations are then incorporated into the evaluation rules for future use. Since our phenomena and categories differ in sizes, we follow Manakhimova et al. (2025) and report aggregate accuracy score percentages on three levels: micro-average (mean over all items, item-weighted), category macro-average, and phenomenon macro-average.

3.6 Manual Annotation Procedure

Manakhimova et al. (2025) evaluate the candidates on standards of basic accuracy and fluency in addition to the targeted phenomenon. However, we decided to follow (Macketanz et al., 2022a) and strictly focused on the targeted phenomenon to assure that the accuracy scores best reflect the models' performance on the given category. In other words, to preserve differences between categories, errors of category-type A should not influence accuracy scores in category B.

3.7 Statistical Analysis

For system comparison, we follow Manakhimova et al. (2025) by first identifying the highest-scoring system and then testing the remaining systems against it using a one-tailed Z-test with $\alpha = 0.05$. Since some of our phenomena contain a low number of test items, we only perform the statistical analysis on an item-weighted category level and micro-average.

4 Results

This section reports a system-level and category-level performance overview of the three systems under investigation: GEMMA 3, LUXMT, and GPT-5 (see Table 1). Due to space constraints, it is not possible to go into all the phenomena and categories in detail. For this reason, we only highlight and illustrate the most notable differences.

4.1 System Performance Overview

Table 1 reports item-weighted average accuracy scores and statistical significance (†) by linguistic category. The results reveal that GPT-5 outperforms the local models in nearly every category, except in *Coordination & Ellipsis*, *Subordination*, and *Named entity & terminology*, where the performance was matched by LUXMT. The local models match ranks with GPT-5 in the three categories mentioned above, and in *LDD & interrogatives* and *Verb valency*. Furthermore, the results indicate improvements of LUXMT over GEMMA 3 in most categories (see Table 2).

4.2 Category Difficulty Analysis

The most challenging categories are *Named entity & terminology* (36.4%), *MWE* (57.4%), *Lexical morphology* (59.1%), *False friends* (60.8%), and *Non-verbal agreement* (62.3%). The results suggest that most difficulties occur on a lexical level.

The least challenging categories include *Coordination & Ellipsis* (91.7%), *Subordination* (91%), and *LDD & interrogatives* (88.9%). These results suggest strong syntactic control.

4.3 Linguistic Analysis

This subsection provides an in-depth linguistic analysis of linguistic phenomena and items that were challenging to MT systems. The analysis reveals the patterns that LUXMT acquired through fine-tuning, improvements over the GEMMA 3 base model, differences between the local models and GPT-5 in their translation performances, as well as strengths and weaknesses that the models have in common. Invalid translations are marked with an asterisk (*).

category	**count**	**Gemma 3**	**LuxMT**	**GPT-5**	**avg**
Ambiguity	50	72.0	74.0	**96.0**†	80.7
Coordination & ellipsis	20	85.0	**95.0**	**95.0**	91.7
False friends	34	52.9	52.9	**76.5**†	60.8
Function word	57	66.7	66.7	**98.2**†	77.2
LDD & interrogatives	30	83.3	90.0	**93.3**	88.9
Lexical morphology	62	50.0	50.0	**77.4**†	59.1
MWE	43	48.8	51.2	**72.1**†	57.4
Named entity & terminology	152	34.2	**37.5**	**37.5**	36.4
Non-verbal agreement	23	43.5	60.9	**82.6**†	62.3
Subordination	37	83.8	**94.6**	**94.6**	91.0
Verb tense/aspect/mood	354	58.8	73.7	**91.8**†	74.8
Verb valency	34	76.5	76.5	**88.2**	80.4
micro-average	896	57.3	65.3	**80.6**†	67.7
macro-average	896	63.0	68.6	**83.6**	71.7

Table 1: Item-weighted average accuracy scores (%) by linguistic category for Gemma 3, LuxMT, and GPT-5. Highest scores per row are shown in bold. Statistical significance based on Z-test is marked by †. Note that the Z-test was not used for categorical macro-average.

4.3.1 Improvements Through Fine-tuning

Table 2 reports item-weighted average category accuracy score differences between LuxMT and Gemma 3, and between GPT-5 and the best performing local model. Comparing LuxMT with the Gemma 3 baseline, the results suggest substantial improvements in *Non-verbal agreement* (+17.4%), *Verb tense, Aspect, Mood* (+14.9%), *Subordination* (+10.8%), and *Coordination & Ellipsis* (+10%).

Regarding the *Non-verbal agreement* category, the largest gain is found in the *Genitive* phenomenon (+21.5%, see Table 4 in Appendix A): a case marker expressing possession. It is important to note that the LB genitive has limited grammatical productivity and can mostly be found in lexicalised phrases similar to MWEs (Döhmer, 2018). For example, in (1), *wéinst menger*, literally ’because of me’, means ’if it’s up to me’ or ’for my sake’. In this sense, its meaning is more hedged and less stern in causality than its literal meaning.

(1) Wéinst menger musse mir keng Äppel kafen.

a. Gemma 3. *Because of my apples, we don’t need to buy any.

b. LuxMT. *Because of me, we don’t have to buy apples.

c. GPT-5. *Because of me we don’t have to buy any apples.

Example (2) shows how LuxMT’s syntactic understanding improved over the Gemma 3 baseline. The test item represents the phenomenon *Gapping* from the category *Coordination & ellipsis* where a verb in the second coordinated clause is omitted:

(2) D’Lena schreift e Bréif an den Tom eng E-Mail.

a. Gemma 3. *Lena is writing a letter to Tom, an email.

b. LuxMT. Lena writes a letter and Tom an email.

Little to no improvements were found in *Named entity & Terminology* (+3.3%), *Ambiguity* (+2%), *Lexical morphology* (+0%), *False friends* (+0%), *Function word* (+0%), and *Verb valency* (+0%). No deteriorations were found on an item-weighted category level. Overall, the results suggest that fine-tuning led to a gain of mostly syntactic and morphological knowledge, and limited lexical knowledge.

4.3.2 Deterioration Through Fine-tuning

Some phenomena show deterioration in accuracy scores of LuxMT in comparison with the Gemma 3 baseline, namely *Noun formation* (-7.1%) and *Idiom* (-10.5%) (see Table 4 in Appendix A). These degradations can be partially attributed to LuxMT translating too literally, as can be seen in Example (3) from the phenomenon *Idiom*. Idioms are multi-word units in which the meaning goes beyond the individual words. Thus, in most cases, idioms cannot be translated literally and have to be translated as a whole.

(3) Chill deng Nippelen!

a. Gemma 3. Cool your jets!

b. LuxMT. *Chill your nipples!

c. GPT-5. Calm down!

category	count	LuxMT – Gemma 3	GPT-5 – Top local
Ambiguity	50	+2.0	+22.0
Coordination & ellipsis	20	+10.0	+0.0
False friends	34	+0.0	+23.6
Function word	57	+0.0	+31.5
LDD & interrogatives	30	+6.7	+3.3
Lexical morphology	62	+0.0	+27.4
MWE	43	+2.4	+20.9
Named entity & terminology	152	+3.3	+0.0
Non-verbal agreement	23	+17.4	+21.7
Subordination	37	+10.8	+0.0
Verb tense/aspect/mood	354	+14.9	+18.1
Verb valency	34	+0.0	+11.7
micro-average	896	+8.0	+15.3
macro-average	896	+5.6	+15.0

Table 2: Performance deltas (%) by linguistic category based on item-weighted average accuracy scores, with category counts.

4.3.3 Performance Gaps between Local Models and GPT-5

The results show that GPT-5 outperforms the local models in every category, except *Coordination & Ellipsis*, *Subordination*, and *Named entity & terminology* where the item-weighted category-level accuracy scores between GPT-5 and LUXMT are even (see Table 2). The biggest performance gaps between GPT-5 and the local models are found in categories *Function word* (+31.5%), *Lexical morphology* (+27.4%), *False friends* (+23.6%), *Ambiguity* (+22%), *Non-verbal agreement* (+21.7%), and *MWE* (+20.9%). These results suggest that GPT-5 has much better overall linguistic abilities in translating LB to EN than the local models, including lexical knowledge and idiomatic expression, morphological and syntactic control.

4.3.4 Common Challenges

The phenomenon with the lowest average accuracy scores across all three models is *Proper name & Location* (18%, see Table 4 in Appendix A). LB place names typically have LB and FR forms, and EN usually borrows from FR. Only more well-known place names were translated correctly. Still, different strategies can be observed: GEMMA 3 and LUXMT guess by using a similar place name of a more known location, while GPT-5 simply leaves the source place name intact. LUXMT did learn some additional place names, hence the model scores higher than GEMMA 3 and GPT-5. Example (4) illustrates these differences:

(4) Ech wunnen zu Märel.

a. GEMMA 3. *I live in Mamer.
b. LUXMT. I live in Merl.
c. GPT-5. *I live in Märel.

The phenomenon *Noun formation* (22.6%) has the second lowest average score. LB examples include verb nominalisation where the stem of the verb is suffixed with the *-er* word formation particle or the *-ert* suffix, which also adds a pejorative connotation, e.g., *Pechert* 'traffic warden' from *pechen* 'to stick'[8]. Models struggle with this construction, as it probably has a low frequency in training corpora. Here, GEMMA 3 and LUXMT interpreted *Pechert* as a bus, while GPT-5 seems to interpret the lexeme as a proper name:

(5) An dëser Strooss passéiert de Pechert zweemol den Dag.

a. GEMMA 3. *The Pechert bus passes this street twice a day.
b. LUXMT. *The bus passes twice a day on this road.
c. GPT-5. *On this street Pechert passes by twice a day.

Another interesting observation is GPT-5's **failure to generalise** linguistic rules, even though there is evidence that the model has enough knowledge to do so. There are multiple examples where GPT-5 correctly translates nouns with the *-ert* suffix. However, Example (6) shows that GPT-5 fails to translate *Schneekert* 'sweet toothed person', despite correctly translating the verb *schneeken* 'to snack [on sweets]'. Thus, the model technically has the required knowledge to generalise and add the noun formation suffix *-ert* to the stem of the

[8] Because the traffic warden 'sticks' the parking fine ticket to the car.

verb *schneeken* to derive the meaning, but it fails to do so:

(6) De Schneekert schneekt gären.
 a. GPT-5. *Mr. Schneekert likes to snack.

Yet, GPT-5 correctly translated the same noun in another sentence (7):

(7) Dee Schneekert géif sech am léifsten zwee Desserte bestellen!
 a. GPT-5. That sweet tooth would most like to order two desserts!

As previous research has already shown (Macketanz et al., 2022b), LLMs struggle with translating idioms. The low average score of 24.6% for the *Idiom* phenomenon corroborates this finding. LLMs tend to translate idioms literally, e.g., in (8) all LLMs returned the same candidate sentence:

(8) Hie kuckt mam rietsen A an déi lénks Täsch.
 a. CANDIDATE. *He looks with his right eye into the left pocket.
 b. SOLUTION. He has a lazy eye.

A characteristic attribute of LB is that the genitive case can also be used in combination with family names, where the genitive marker becomes assimilated with the family name. LLMs tend to struggle with this, mainly because they avoid altering named entities, e.g., in (9):

(9) Mir gi mat Müllesch iessen.
 a. GEMMA 3. *We are going to eat with Müllesch.
 b. LUXMT. *We go with Müllesch to eat.
 c. GPT-5. *We are going to eat with Müllesch.
 d. SOLUTION. We are going to eat out with the Müllers.

In line with previous research (Avramidis et al., 2020; Manakhimova et al., 2025) *Resultative predicates* are challenging for the three models. This phenomenon includes constructions where the adjective describes the result of an action expressed by a verb. Similar to MWEs, these are often language-specific and literal translations risk leading to errors. For example, in (10), the correct translation for *eidel drénken* is ’to empty’ or ’to finish’. A literal translation like ’to drink empty’ is a mistranslation. LLMs also fail to use the correct adjective, e.g., in (11) *platt lafen* ’to trample down’ instead of ’to trample flat’ or ’to run flat’.

(10) Si huet d’Taass eidel gedronk.
 a. GEMMA 3. *He/She drank the tea empty.
 b. LUXMT. *She drank the cup empty.
 c. GPT-5. *She drank the cup empty.

(11) D’Joggeren hunn d’Wiss platt gelaf.
 a. GEMMA 3. *The joggers ran over the meadows.
 b. LUXMT. *The joggers ran flat on the ground.
 c. GPT-5. *The joggers have trampled the meadow flat.

Another noteworthy issue concerns mistranslations that could be attributed to false friends in non-target languages. These errors were annotated but not factored into the accuracy scores. For example, it is plausible that the model interpreted the LB lexeme *Wiss* ’lawn’ as a false friend of DE *Wissen* ’knowledge’, and LB *geméit* ’mown’ or ’mowed’ as DE *gemessen* ’measured’:

(12) D’Wiss gëtt ëmmer mëttwochs geméit.
 a. GEMMA 3. The lawn is always mowed on Wednesdays.
 b. LUXMT. The knowledge is always measured on Wednesdays.

5 Discussion

What Luxembourgish linguistic knowledge do local LLMs acquire through fine-tuning? The Lux-MT-Test-Suite shows that fine-tuning local LLMs improves their performance in a wide range of linguistic categories and phenomena, reflecting increased lexical, morphological, and syntactical knowledge and control. However, the largest gains are restricted to morphological and syntactical knowledge, while the gain in lexical knowledge is limited. It is also observable that the fine-tuned model’s improvements do not always stack on top of the base model: the overall performance may increase, but the fine-tuned model may mistranslate items that the base model translates correctly (see Table 4 in Appendix A). It is unclear whether an underlying issue, such as catastrophic forgetting

(Liu and Niehues, 2025), or merely inconsistent translation performance is the cause. Iterating the model multiple times over the same test suite could provide some clarity.

What are the performance gaps between local LLMs and SOTA proprietary LLMs? The fine-grained evaluation demonstrated that, in a low-resource setting, GPT-5 is still ahead of local LLMs, even when local LLMs are fine-tuned. The difference between the best performing local model and GPT-5 was statistically significant in five out of 12 categories.

The test suite also helped identify common challenges that still persist. LLMs often translate too literally, resulting in translationese (Gellerstam, 1986), and struggle with idiomatic language, generating text that sounds characteristically non-human. The test suite also showed that, in a low-resource case, lexical LLMs have gaps in lexical knowledge. These cases also revealed that different LLMs have different strategies, resulting in different error-profiles, e.g., interpreting unknown nouns as named entities or relying on similarities with another language. These findings are largely in-line with previous research (Manakhimova et al., 2025).

6 Conclusion

To conclude, this paper introduced the Lux-MT-Test-Suite enabling a fine-grained evaluation of LB-EN MT. To showcase the test suite, we compared a popular local LLM (GEMMA 3) with a fine-tuned counterpart (LUXMT) and a proprietary SOTA LLM (GPT-5), identifying improvements achieved through fine-tuning, performance gaps between local and proprietary models, and differences in translation strategies. The fine-tuned model improved over its baseline across a wide range of categories with minor degradation in some phenomenon-specific performance. The fine-tuned model matched GPT-5's model performance in various categories, but some performance gaps between LUXMT and GPT-5 remain. Common challenges include idiomatic expressions and lexical knowledge, and LLMs still frequently translate too literally, resulting in translationese.

7 Limitations

Annotation: No inter-annotator agreement score could be calculated as this study relied on a single annotator. The results should be interpreted cautiously.

Purpose: To ensure the accuracy of the phenomenon and category-level scores, the candidates were evaluated only on the phenomenon of interest as much as possible. This means that many errors had to be ignored. Consequently, the accuracy scores serve to diagnose translation difficulties rather than to reflect translation quality, as the phenomena are not indicative of the severity of translation errors.

Score reliability: Since some of the data is publicly available, such as the LOD examples, it is possible that the data was scraped and unintentionally included in the training set of the GEMMA 3 base model and GPT-5. This possible contamination risks inflating model performance (Sainz et al., 2023). Another limitation is the binary evaluation of correctness. While binary evaluations are useful for quantifying quality, it is ultimately reductive, especially when concerned with human language which is inherently nuanced and ambiguous (Manakhimova et al., 2025). Moreover, a high score on a phenomenon or category should not necessarily be taken as a guarantee that the MT system masters the given phenomenon or category, as it may be that the test items are not difficult enough (Isabelle et al., 2017).

Distribution: The number of test items per category or phenomenon is not representative of any distribution in corpora or real-world settings. This means that a model performing better on average is not necessarily the most performant for any application. Furthermore, some phenomena and categories contain a small amount of test items, leading to potentially skewed accuracy scores.

Sentence-level evaluation: The Lux-MT-Test-Suite evaluates performance on a segment-level, leaving translation difficulties that might arise on a paragraph or discourse level unexplored (Manakhimova et al., 2025). Future research could include test items that target paragraph-level phenomena.

Future development of the test suite will increase the number of items, include an EN-LB translation direction and implement more LB-specific phenomena.

Acknowledgments

This research is part of the Lux-ASR 2 project[9], supported by the Chambre des Députés.

I would like to sincerely thank Anne-Marie Lutgen, Dr. Alistair Plum, and Dr. Nina Hosseini-

[9] https://infolux.uni.lu/lux-asr/

Kivanani for their valuable feedback, as well as my supervisor, Prof. Dr. Peter Gilles, for his guidance during the project.

References

Dimitra Anastasiou. 2022. Enrich4all: A first luxembourgish bert model for a multilingual chatbot. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 207–212.

Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. A test suite for the evaluation of portuguese-english machine translation. In *International Conference on Computational Processing of the Portuguese Language*, pages 15–25. Springer.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. *arXiv preprint arXiv:2010.06359*.

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

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Caroline Döhmer. 2018. A new perspective on the luxembourgish genitive. In *Germanic genitives*, pages 15–36. John Benjamins Publishing Company.

Caroline Döhmer. 2020. *Aspekte der luxemburgischen Syntax*. BoD–Books on Demand.

Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. Schnëssen. surveying language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7(s1):20190031.

Fernand Fehlen and Andreas Heinz. 2016. *Die Luxemburger Mehrsprachigkeit: Ergebnisse einer Volkszählung*. transcript Verlag.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Daniela Gierschek. 2022. Detection of sentiment in luxembourgish user comments.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina HOSSEINI KIVANANI. 2023a. Asrlux: Automatic speech recognition for the low-resource language luxembourgish. In *20. International Conference of Phonetic Sciences (ICPhS)*. Guarant International, Prague, Unknown/unspecified.

Peter Gilles, Nina HOSSEINI KIVANANI, and Léopold Edem Ayité Hillah. 2023b. Lux-asr: Building an asr system for the luxembourgish language. In *2022 IEEE Spoken Language Technology Workshop (SLT) SLT 2022*.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Ulrich Heid and Elke Hildenbrand. 1991. Some practical experience with the use of test suites for the evaluation of systran. In *the Proceedings of the Evaluators' Forum, Les Rasses. Citeseer*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, Suzhou, China. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, and 1 others. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Conditions for catastrophic forgetting in multilingual translation. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 347–359.

Arle Lommel, Serge Gladkoff, Alan K Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, and 1 others. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. Luxembert: Simple and practical data augmentation in language model pre-training for luxembourgish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5080–5089.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 115–127.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022a. A linguistically motivated test suite to semi-automatically evaluate german–english machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022b. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shushen Manakhimova, Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski, and Eleftherios Avramidis. 2025. Fine-grained evaluation of english-russian mt in 2025: Linguistic challenges mirroring human translator training. In *Proceedings of the Tenth Conference on Machine Translation*, pages 866–877.

Andrianos Michail, Corina Raclé, Juri Opitz, and Simon Clematide. 2025. Adapting multilingual embedding models to historical luxembourgish. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 291–298.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*.

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.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. Luxbank: The first universal dependency treebank for luxembourgish. *arXiv preprint arXiv:2411.04813*.

Alistair Plum, Tharindu Ranasinghe, and Christoph Purschke. 2025. Text generation models for luxembourgish with limited data: A balanced multilingual strategy. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 93–104.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. *arXiv preprint arXiv:2506.19571*.

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or hold? automatic comment moderation in luxembourgish news articles. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 968–978.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Yewei Song, Saad Ezzini, Jacques Klein, Tegawende Bissyande, Clément Lefebvre, and Anne Goujon. 2023. Letz translate: Low-resource machine translation for luxembourgish. In *2023 5th International Conference on Natural Language Processing (ICNLP)*, pages 165–170. IEEE.

Yewei Song, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2025. Is llm the silver bullet to low-resource languages machine translation? *arXiv preprint arXiv:2503.24102*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480. European Association for Machine Translation.

Julian Valline, Cedric Lothritz, and Jordi Cabot. 2025. Luxit: A luxembourgish instruction tuning dataset from monolingual seed data. *arXiv preprint arXiv:2510.24434*.

Andrew Way. 1991. Developer-oriented evaluation of mt systems. In *Proceedings of the evaluators' forum*, pages 237–244.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Phenomenon-level Overview

phenomenon	count	Gemma 3	LuxMT	GPT-5	avg
Ambiguity	50	72.0	74.0	**96.0**†	80.7
Lexical ambiguity	50	72.0	74.0	**96.0**	80.7
Coordination & ellipsis	20	85.0	**95.0**	**95.0**	91.7
Gapping	10	70.0	**90.0**	**90.0**	83.3
Sluicing	10	**100.0**	**100.0**	**100.0**	100.0
False friends	34	52.9	52.9	**76.5**†	60.8
Function word	57	66.7	66.7	**98.2**†	77.2
Focus particle	40	55.0	55.0	**97.5**	69.2
Question tag	17	94.1	94.1	**100.0**	96.1
LDD & interrogatives	30	83.3	90.0	**93.4**	88.9
Multiple connectors	9	88.9	88.9	**100.0**	92.6
Topicalization	5	60.0	**80.0**	**80.0**	73.3
Wh-movement	16	87.5	**93.8**	**93.8**	91.7
Lexical morphology	62	50.0	50.0	**77.5**†	59.1
Gender	34	82.4	88.2	**97.1**	89.2
Noun formation	28	10.7	3.6	**53.6**	22.6
MWE	43	48.8	51.2	**72.1**†	57.4
Collocation	12	75.0	**91.7**	**91.7**	86.1
Idiom	19	15.8	5.3	**52.6**	24.6
Prepositional MWE	5	60.0	**80.0**	60.0	66.7
Verbal MWE	7	85.7	85.7	**100.0**	90.5
Named entity & terminology	152	34.2	**37.5**	**37.5**	36.4
Date	9	**100.0**	**100.0**	**100.0**	100.0
Proper name & location	100	20.0	**22.0**	12.0	18.0
Festivities	43	53.5	60.5	**83.7**	65.9
Non-verbal agreement	23	43.5	60.9	**82.6**†	62.3
Coreference	9	77.8	**88.9**	**88.9**	85.2
Genitive	14	21.4	42.9	**78.6**	47.6
Subordination	37	83.8	**94.6**	**94.6**	91.0
Adverbial clause	8	**87.5**	**87.5**	**87.5**	87.5
Cleft sentence	5	80.0	**100.0**	**100.0**	93.3
Infinitive clause	10	80.0	90.0	**100.0**	90.0
Object clause	6	83.3	**100.0**	**100.0**	94.4
Subject clause	8	87.5	**100.0**	87.5	91.7
Verb tense/aspect/mood	354	58.8	73.7	**91.8**†	74.8
Conditional	8	62.5	75.0	**87.5**	75.0
Ditransitive	84	63.1	72.6	**91.7**	75.8
Gerund	9	66.7	66.7	**88.9**	74.1
Imperative	10	70.0	**90.0**	**90.0**	83.3
Intransitive	75	61.3	78.7	**96.0**	78.7
Reflexive	84	42.9	57.1	**90.5**	63.5
Transitive	84	65.5	85.7	**90.5**	80.6
Verb valency	34	76.5	76.5	**88.2**	77.1
Case government	9	88.9	88.9	**100.0**	52.4
Mediopassive voice	8	**75.0**	**75.0**	**75.0**	75.0
Passive voice	10	90.0	90.0	**100.0**	90.0
Resultative predicates	7	42.9	42.9	**71.4**	92.6
micro-average	896	57.3	65.3	**80.6**†	67.6
phen. macro-average	896	67.5	74.9	**86.3**	76.2
categ. macro-average	896	63.0	68.6	**83.6**	71.4

Table 3: Phenomenon-level accuracy scores (%) for GEMMA 3, LUXMT, and GPT-5. Highest scores per row are shown in bold. Statistical significance based on Z-test is marked by †. Note that no Z-test was used for macro-averages.

phenomenon	**count**	**LuxMT – Gemma 3**	**GPT-5 – Top local**
Ambiguity	50	+2.0	+22.0
Lexical ambiguity	50	+2.0	+22.0
Coordination & ellipsis	20	+10.0	+0.0
Gapping	10	+20.0	+0.0
Sluicing	10	+0.0	+0.0
False friends	34	+0.0	+23.6
Function word	57	+0.0	+31.5
Focus particle	40	+0.0	+42.5
Question tag	17	+0.0	+5.9
LDD & interrogatives	30	+6.7	+3.4
Multiple connectors	9	+0.0	+11.1
Topicalization	5	+20.0	+0.0
Wh-movement	16	+6.3	+0.0
Lexical morphology	62	+0.0	+27.5
Gender	34	+5.8	+8.9
Noun formation	28	-7.1	+42.9
MWE	43	+2.4	+20.9
Collocation	12	+16.7	+0.0
Idiom	19	-10.5	+36.8
Prepositional MWE	5	+20.0	-20.0
Verbal MWE	7	+0.0	+14.3
Named entity & terminology	152	+3.3	+0.0
Date	9	+0.0	+0.0
Proper name & location	100	+2.0	-10.0
Festivities	43	+7.0	+23.2
Non-verbal agreement	23	+17.4	+21.7
Coreference	9	+11.1	+0.0
Genitive	14	+21.5	+35.7
Subordination	37	+10.8	+0.0
Adverbial clause	8	+0.0	+0.0
Cleft sentence	5	+20.0	+0.0
Infinitive clause	10	+10.0	+10.0
Object clause	6	+16.7	+0.0
Subject clause	8	+12.5	-12.5
Verb tense/aspect/mood	354	+14.9	+18.1
Conditional	8	+12.5	+12.5
Ditransitive	84	+9.5	+19.1
Gerund	9	+0.0	+22.2
Imperative	10	+20.0	+0.0
Intransitive	75	+17.4	+17.3
Reflexive	84	+14.2	+33.4
Transitive	84	+20.2	+4.8
Verb valency	34	+0.0	+11.7
Case government	9	+0.0	+11.1
Mediopassive voice	8	+0.0	+0.0
Passive voice	10	+0.0	+10.0
Resultative predicates	7	+0.0	+28.5
micro-average	896	+8.0	+15.3
phen. macro-average	896	+7.4	+11.4
categ. macro-average	896	+5.6	+15.0

Table 4: Performance deltas (%) by linguistic phenomenon, with counts.