Translation Asymmetry in LLMs as a Data Augmentation Factor: A Case Study for 6 Romansh Language Varieties

Jannis Vamvas[1] **Ignacio Pérez Prat**[2] **Angela Heldstab**[1]
Dominic P. Fischer[1] **Sina Ahmadi**[1] **Rico Sennrich**[1]

[1]University of Zurich [2]Lia Rumantscha

Correspondence: vamvas@cl.uzh.ch, ignacio.perez.prat@rumantsch.ch

Abstract

Recent strategies for low-resource machine translation rely on LLMs to generate synthetic data based on text in higher-resource languages. We revisit this idea for Romansh, a language with 6 distinct varieties. LLMs tend to confuse these varieties when translating *into* Romansh, but they are quite good at translating *out of* Romansh into a high-resource language such as German. Due to this asymmetry, the direction of data augmentation is a crucial choice. We find that contrary to recent strategies, creating synthetic translations into the *higher*-resource language is the superior approach, and only this approach allows us to surpass a Gemini 3 Pro baseline on German→Romansh translation (+23 BLEU over Gemini in the lowest-resource variety). A human evaluation confirms that our experiments yield the first model that generates fluent translations in the individual Romansh varieties.

🤗 Fine-tuned model	`hf.co/ZurichNLP/romansh-nllb-1.3b-ct2`
🤗 Back-translations	`hf.co/datasets/ZurichNLP/romansh-backtranslated`
🤗 Human ratings	`hf.co/datasets/ZurichNLP/romansh-mt-evaluation`
Evaluation code	`github.com/ZurichNLP/romansh_mt_eval`
Production system	`alas.liarumantscha.ch`

1 Introduction

Traditionally, both forward translation and back-translation have been found to be helpful data augmentation strategies for neural MT (see, e.g., Burlot and Yvon, 2018), and the choice of strategy depended on the available monolingual data. In recent work, large-scale LLM-based forward translation from English was used to create synthetic training data for low-resource MT (de Gibert et al., 2025). While this may be effective for some language pairs, we draw attention to the inherent asymmetry of LLM capabilities for low-resource or multi-variety languages (Figure 1).

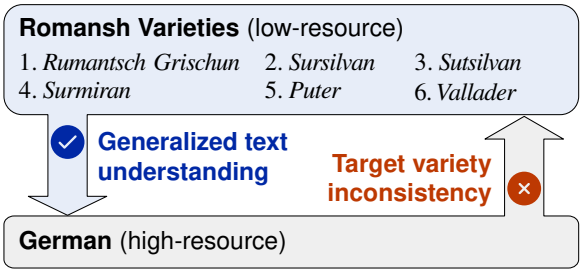

Figure 1: LLMs have asymmetric translation capabilities regarding low-resource or multi-variety languages like Romansh. In the case of Romansh, they demonstrate a general understanding of all varieties when translating *out of* the language, but they fail to adhere to a specific target variety when translating *into* the language. This asymmetry is relevant for data augmentation.

This paper studies Romansh, a minority language in Switzerland spoken by 40–60,000 speakers across several Alpine valleys. Since Romansh has 6 distinct varieties, which are often mutually unintelligible, the development of variety-aware MT is a strong community need. Romansh is also a prime example of translation asymmetry: LLMs are known to excel at understanding Romansh text in any variety when translating *out of* Romansh, but struggle to produce text in the desired variety when translating *into* Romansh (Vamvas et al., 2025; Project Apertus et al., 2026). Therefore, generating synthetic Romansh text with LLMs might not be the optimal approach for data augmentation.

We present a controlled German↔Romansh case study that empirically compares the two directions of data augmentation: Either starting with text in the high-resource language and generating synthetic translations in the low-resource language (***HR→LR augmentation***), or investing into manual data collection in the low-resource language

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 4837–4861
October 24-29, 2026 ©2026 Association for Computational Linguistics

Rumantsch Grischun (RG)	*Quai n'hai jau betg ditg.*
Sursilvan	*Quei hai jeu buca detg.*
Sutsilvan	*Quegl ve jou betga getg.*
Surmiran	*Chegl vaia betg detg.*
Puter	*Que nu d'heja dit.*
Vallader	*Quai nun haja dit.*
[German]	*Das habe ich nicht gesagt.*
[English]	*I didn't say that.*

Figure 2: The varieties of Romansh are highly diverse, as shown in this example from the WMT24++ benchmark (Deutsch et al., 2025; Vamvas et al., 2025).

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	39.8	18.1	8.7	11.9	11.9	15.4
Surs.	25.2	29.5	8.4	10.4	9.7	12.2
Suts.	27.2	17.5	10.0	12.2	11.0	13.5
Surm.	19.6	13.0	9.3	17.8	10.7	11.7
Puter	14.1	10.0	7.4	8.7	23.1	23.3
Vall.	16.2	11.1	7.4	8.6	20.5	27.0

(a) Gemini 2.5 Flash

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	48.4	19.9	8.6	12.2	12.4	15.5
Surs.	19.0	44.5	8.3	9.4	8.2	9.8
Suts.	9.0	9.3	40.5	10.1	6.7	7.5
Surm.	12.5	10.5	9.7	43.0	8.1	8.9
Puter	11.9	8.9	6.3	8.0	44.9	25.0
Vall.	14.8	10.1	6.7	8.1	24.7	44.6

(b) Our system

Figure 3: When translating from German into Romansh, Gemini 2.5 Flash often produces the wrong target variety, which can be visualized through a confusion matrix that evaluates the system translations against the references for the individual varieties. Our fine-tuned NLLB model adheres to the target varieties and achieves higher BLEU.

and then generating synthetic translations in the high-resource language (***LR→HR augmentation***).

To compare these augmentation directions, we fine-tune NLLB (Costa-jussà et al., 2024) on the augmented data and evaluate German↔Romansh translation across the six varieties. Surprisingly, we find that LR→HR augmentation is the better approach for both downstream translation directions (German→Romansh *and* Romansh→German). Thus, for German→Romansh MT, back-translating low-resource text outperforms forward-translating high-resource text. For Romansh→German MT, forward-translating low-resource text is likewise superior to back-translating high-resource text.

Our case study suggests that authentic data in low-resource languages remains highly beneficial for LLM-based data augmentation, and therefore that collecting monolingual text is still an important building block of low-resource MT. In summary, our main contributions are:

- We provide a controlled comparison of LLM-based data augmentation in a low-resource scenario with 6 distinct target varieties.
- We validate prompting strategies such as few-shot prompting and dictionary prompting for data augmentation.
- We release our fine-tuned MT models, the synthetic back-translations used for LR→HR augmentation, 10k+ human quality ratings, and the evaluation code.

2 Background: Language Varieties of Romansh

Romansh, one of the four official languages of Switzerland, is a minority language spoken in only one Swiss canton (Grisons) and is considered endangered. It has been argued that it is inaccurate to call Romansh a single language (Caviezel, 1993) and that it is instead a continuum with no overarching language variety (Liver, 2010; Grünert, 2024). Within Grisons, the Romansh-speaking regions can be separated into five territories, where each of the five has their own written tradition, called 'idiom'. An idiom is thus a standardized version of Romansh language varieties spoken within one of the Romansh-speaking regions, with its own orthography (see Figure 2) and its own literary tradition. This split standardization dates back to the 1500s with religious and literary publications and was emphasized due to the lack of a centralized and common cultural space (Grünert, 2024).

The largest speaker groups are Sursilvan (est. 50% of Romansh speakers) and Vallader (20%), while Sutsilvan has very few speakers (3%). The other two are Surmiran (10%) and Puter (12%) (Furer, 2005). This distribution is broadly reflected in the available data (see Appendices F and H).

Rumantsch Grischun (RG), unlike the other five varieties, is not an idiom that was standardized based on spoken language, but rather a constructed language mostly based on Sursilvan, Surmiran, and Vallader (Liver, 2010). It was intended for govern-

mental and institutional communication and explicitly introduced as complementary to the existing idioms (Coray, 2008). The canton and the federal government use RG for official publications and communications. Most web data readily available is in RG, and for this reason, prior work on MT for Romansh has focused on RG only (e.g., Müller et al., 2020; Kudugunta et al., 2023, as well as a commercial system by Supertext). However, Romansh speakers use their native idioms both in school and at work, and therefore, most speakers have passive knowledge of RG at best. Variety-aware MT would allow Romansh speakers a greater degree of access to content produced in other languages or other Romansh varieties.

3 LLM Translation Asymmetry

It is well known that LLMs perform better with low-resource languages as a source than as a target (e.g., Bawden and Yvon, 2023; Zhu et al., 2024; Omnilingual MT Team et al., 2026). Recent evaluations have shown the same asymmetry for Romansh (Vamvas et al., 2025; Project Apertus et al., 2026), and Table 1 reproduces this finding with four LLMs.

There are several potential explanations for the asymmetry. One explanation is *transfer asymmetry*: Knowledge transfer from related, higher-resource languages may compensate for limited low-resource language exposure more effectively for understanding, rather than generation. An alternative explanation is *standardization asymmetry*: Some low-resource languages might have more diverse variation than high-resource languages (relative to their representation in the training data), which could make the task of translating into these languages more ambiguous for a general-purpose LLM. Finally, *tokenization disparity* could mean that the generation of text in low-resource languages requires more decisions and that the errors and biases of autoregressive architectures (Lesci et al., 2025) compound more easily.

While it remains open to what degree these speculative explanations apply to the German↔Romansh language pair, we assume that *standardization asymmetry* is especially important. The confusion matrix in Figure 3a illustrates how LLMs struggle to produce a specific variety of Romansh on demand: when translating text from German into specific varieties such as Sutsilvan, Surmiran or Puter, Gemini 2.5 Flash achieves a *lower* BLEU score under the reference translations in the desired variety than under the reference translations in other, higher-resource varieties. Some more recent versions of Gemini are less confused, but still strongly favor higher-resource varieties, especially RG (Appendix L).

In this paper, we show that LLM translation asymmetry has direct implications for data augmentation.

4 Data Augmentation Methods

4.1 Forward and Back-translation

Forward translation generates synthetic targets from monolingual source text, while *back-translation* generates synthetic sources from monolingual target text. Comparative studies have found both methods effective for NMT (Burlot and Yvon, 2018; Bogoychev and Sennrich, 2020), and both remain widely used (Kocmi et al., 2025). Recently, de Gibert et al. (2025) found that large-scale forward translation with GPT-4o improves NMT for 7 low-resource languages, such as Basque and Georgian. Frontull and Moser (2024) combined forward and back-translation in an iterative process for Italian↔Ladin with GPT 3.5 back-translation as a seed. However, there exists no systematic comparison of LLM-based forward and back-translation.

In this paper, we perform a comparison in a controlled setting for German↔Romansh MT, fixing the LLM (Gemini 2.5 Flash), prompt, and downstream NMT model (NLLB). We train our model variants jointly in both directions and on all varieties, using each sample bidirectionally. While this multilingual setting differs from traditional bilingual back-translation (Sennrich et al., 2016), it is more in line with our production requirements.

As a consequence, all synthetic translations can be considered both forward translations and back-translations, depending on the direction in which the downstream model will be used, but the open question is whether they should be created in the low-resource language (*HR→LR augmentation*, like de Gibert et al., 2025), or whether they should be created in the high-resource language based on text in the low-resource language (*LR→HR augmentation*). We will use these two terms in the next sections to avoid any ambiguity.

4.2 Dictionary Prompting

We experiment with a variant of LLM-based augmentation that appends bilingual dictionary entries

to the prompt (Ghazvininejad et al., 2023; Court and Elsner, 2024). Due to the high cost of this style of prompting, we test it only in the LR→HR augmentation direction.

The dictionary information consists of the Romansh lemma for each word in the input document, together with potential German word-level translations and morphological analyses of the words, where available. For example, the Sursilvan word *stos* is explained to the LLM as follows:

```
stos (lemma: stuer) → müssen
[PoS=V;Mood=IND;Tense=PRS;Person=2;Number=SG]
```

Appendix K presents a complete example prompt.

5 Experimental Setup

5.1 Evaluation Data

For evaluation, we primarily use the Romansh version of the WMT24++ benchmark (Vamvas et al., 2025), paired with the German version from Deutsch et al. (2025). This benchmark contains 998 text segments from 4 domains (literary, news, social, speech). We split the benchmark into a validation split (first 50% documents per domain) and a test split (remaining 50%), to allow for hyperparameter optimization and prompt engineering based on the validation split. The final results are based on the test split.

We additionally evaluate on the more recent BOUQuET benchmark (Andrews et al., 2025), which provides 854 German–Romansh sentence pairs covering 8 domains and combinations of 12 linguistic registers. This benchmark is only available for the RG variety of Romansh.

5.2 LLM-based Data Augmentation

For creating the synthetic training data, we prompt the LLM with the instruction *"Translate the following text into {language},"* translating segments of up to 500 tokens in a single request. The full prompts are reported in Appendix J. We use greedy decoding for the data augmentation, leaving sampling-based augmentation (Edunov et al., 2018) to future work.

Furthermore, results obtained on our validation split of WMT24++ (Table 1) lead us to make the following choices:

- **Gemini 2.5 Flash** (Comanici et al., 2025) achieves the best scores in both translation directions compared to three other LLMs of similar cost, confirming prior findings by Vamvas et al. (2025). Thus, we use this model for further experiments.

- **Disabling reasoning** allows us to save cost for the large-scale data augmentation experiments. Enabling reasoning (with a budget of 2048 tokens) only leads to modest gains on the validation set (+0.5 BLEU for German→Romansh; +0.1 COMET for Romansh→German).

- **Few-shot examples** are crucial for instruction following (+3.2 BLEU for German→Romansh; +5.8 COMET for Romansh→German). Zero-shot prompting without any examples results in translations into English instead of German, leading to a very poor BLEU score of 3.2.

- **Dictionary prompting** shows a positive effect (+0.4 COMET). Even though the effect is relatively small in terms of COMET, we still include this as a variant in the large-scale data augmentation experiment and the human evaluation, since the technique might affect long-tail phenomena that are harder to quantify with automatic metrics.

In total, we run four large-scale data augmentation experiments with Gemini 2.5 Flash (different data mixes, with or without dictionary prompting), with a total cost of approximately 2,400 USD (Appendix G).

	DE→RM	RM→DE
Approach	**BLEU**	**BLEU / COMET**
Baseline LLMs		
Llama 3.3 (70B)	11.1	27.1 / 70.1
Apertus (70B)	13.7	35.6 / 84.3
GPT-4o	22.7	48.3 / 89.4
Gemini 2.5 Flash		
Baseline prompt (J.1)	23.6	51.2 / 91.0
– with reasoning	24.1	50.1 / 91.1
– without few-shot examples	20.4	3.2 / 85.2
Dictionary prompting (J.2)	-	51.4 / 91.4
– with reasoning	-	49.4 / 91.5

Table 1: Validation results for multiple LLMs and data augmentation strategies, based on our validation split of the WMT24++ benchmark and averaged across the six varieties of Romansh. The results for baseline LLMs are reported for comparison, and only Gemini 2.5 Flash is used for the experiments below.

Variety	# Raters	# Fluency ratings	# Document accuracy ratings	# Segment accuracy preferences
RG	2	1,117 (51 overlap)	168 (16 overlap)	531 (159 overlap)
Sursilvan	5	1,015 (64 overlap)	204 (21 overlap)	684 (271 overlap)
Sutsilvan	3	987 (139 overlap)	166 (37 overlap)	704 (341 overlap)
Surmiran	2	1,017 (238 overlap)	168 (23 overlap)	668 (303 overlap)
Puter	3	929 (160 overlap)	158 (0 overlap)	425 (0 overlap)
Vallader	2	827 (0 overlap)	166 (0 overlap)	484 (0 overlap)
Total	**17**	**5,892**	**1,030**	**3,496**

Table 2: Human evaluation statistics: number of raters and annotations by target variety. The overlap counts in parentheses refer to items annotated by more than one rater, which we use for inter-rater agreement analysis.

Metric	Fluency	Document accuracy	Segment accuracy
Item-level Pearson correlation between raters	67.6	31.7	42.3
Item-level pairwise agreement rate between raters	68.5	59.0	57.8
System-level Spearman correlation between raters	92.0	75.0	90.0

Table 3: Inter-rater agreement metrics for the human evaluation, averaged across the 4–5 target varieties in which overlapping annotations have been collected.

5.3 Training Data

As a prerequisite for the experiment, we compile monolingual and parallel data for the Romansh language varieties (Appendix F).

For LR→HR augmentation, we use 140M tokens of authentic monolingual Romansh data, mostly from web corpora, news and schoolbooks (Table 10). In total, we include data from 12 different sources, many of which are already publicly available, or have been made available by us as part of this publication to enable future reproductions.

For HR→LR augmentation from German, we compile two different datasets of an equivalent size:

- **Europarl:** In this variant, we follow de Gibert et al. (2025) and sample from German Europarl (Koehn, 2005). This approach is directly comparable to prior work but is based on text in a different domain than the authentic Romansh data used for LR→HR augmentation, which could confound the comparison.
- **Newscrawl / FineWeb2:** This variant uses a German data mix that is comparable to the authentic Romansh data: 50% news text via Newscrawl 2023 (Kocmi et al., 2025) and 50% web text via FineWeb2 (Penedo et al., 2025). Figure 5 shows that this data mix closely approximates the domain composition used for LR→HR augmentation.

In all cases, we add 23M tokens of authentic parallel data, including around 3M word forms from Romansh–German bilingual dictionaries. The LLM generates synthetic translations at the document level, which we then split into smaller segments for NMT fine-tuning. We re-label the variety of all Romansh segments (including synthetic text) using an LID classifier for Romansh varieties (Model et al., 2026). Further preprocessing details and final training data statistics are in Appendices C and H, showing that comparable amounts of training data are used for all approaches.

5.4 NMT Fine-tuning

We fine-tune the NLLB-200-Distilled 1.3B model (Costa-jussà et al., 2024), a multilingual NMT model with a transformer-based encoder-decoder architecture. Since NLLB does not include Romansh as a pre-trained language, we extend the model vocabulary with 6 new language tokens corresponding to each Romansh variety. We use temperature sampling with $T = 1.5$ to upsample underrepresented translation directions. Other hyperparameters are reported in Appendix A.

5.5 Automatic Evaluation

For German→Romansh, we compute BLEU scores (Papineni et al., 2002) using SacreBLEU (Post, 2018)[1], due to the lack of a trained metric for this target language (complemented with a human evaluation, see Section 5.6 below). For Romansh→German, we primarily report xCOMET (Guerreiro et al., 2024), replicating the setup of Vamvas et al. (2025). We ap-

[1]Version: `#:1|c:mixed|e:no|tok:13a|s:exp|v:2.5.1`

System	WMT24++						BOUQuET	Avg.
	RG	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader	RG	
Gemini 2.5 Flash	40.3	28.0	10.6	16.9	22.0	27.3	27.2	24.6
Gemini 3 Pro (preview)	45.3	37.1	17.3	27.1	34.7	36.1	31.3	32.7
Gemini 3.7 Flash	40.9	34.2	14.8	22.1	30.4	29.8	29.8	28.9
Fine-tuned NLLB								
No data augmentation	35.9	29.4	29.2	25.3	31.0	26.0	27.9	29.2
HR→LR augmentation:								
– Europarl data	40.8	31.3	12.4	18.7	23.8	30.7	29.6	26.8
– Newscrawl / FineWeb2	41.4	30.4	12.1	18.5	24.6	29.4	29.7	26.6
LR→HR augmentation:								
– Baseline prompt	48.1	44.3	**40.6**	42.3	44.8	44.3	**33.1**	42.5
– Dictionary prompting	**48.4**	**44.5**	40.5	**43.0**	**44.9**	**44.6**	32.6	**42.6**

Table 4: **BLEU scores for German→Romansh** translation on the second half of the WMT24++ benchmark, together with BOUQuET results for Rumantsch Grischun. Bold denotes the best system; underline indicates the best variant among the fine-tuned NLLB models; the last column reports the arithmetic mean of the seven columns.

System	WMT24++						BOUQuET	Avg.
	RG	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader	RG	
Gemini 2.5 Flash	93.7	93.1	89.8	91.7	92.3	92.7	97.6	93.0
Gemini 3 Pro (preview)	93.8	93.8	92.5	93.4	93.0	93.8	97.8	94.0
Gemini 3.7 Flash	**94.3**	**94.0**	**93.3**	**93.9**	**93.4**	**94.1**	**97.9**	**94.4**
Fine-tuned NLLB								
No data augmentation	81.9	80.9	80.2	79.1	79.0	79.3	93.3	82.0
HR→LR augmentation:								
– Europarl data	89.1	89.6	88.6	88.6	88.9	88.4	96.2	89.9
– Newscrawl / FineWeb2	91.7	90.9	89.8	89.5	90.7	90.1	97.0	91.4
LR→HR augmentation:								
– Baseline prompt	92.4	92.3	90.8	91.1	91.6	91.3	97.3	92.4
– Dictionary prompting	92.5	92.4	91.1	91.4	91.6	91.5	97.3	92.5

Table 5: **COMET scores for Romansh→German** translation on the second half of the WMT24++ benchmark, together with BOUQuET results for Rumantsch Grischun. Bold denotes the best system; underline indicates the best variant among the fine-tuned NLLB models; the last column reports the arithmetic mean of the seven columns.

ply the model version `XCOMET-XL`[2] in *reference-only* mode, providing only the German translation and reference without the Romansh source, as Romansh was not included in xCOMET's training data. Following Kocmi et al. (2024), xCOMET scores are macro-averaged across domains to account for varying segment granularities.

5.6 Human Evaluation

We recruited 18 native speakers of all six Romansh varieties to blindly compare the three best systems in the German→Romansh direction, plus the human references. Fluency was rated on the level of individual segments, while accuracy was rated on the document level; both on a 7-point scale.

For most varieties, a part of the items were rated by more than one rater (Table 2). This allows us to calculate inter-rater agreement metrics (Table 3 reports the numbers on average across target varieties, and Appendix N.2 provides a per-variety analysis.) System-level Spearman correlations mostly range from 80 to 100 for both fluency and accuracy, while item-level Pearson correlations are moderate (52–82 for fluency, 17–50 for accuracy). [3]

To calculate system scores from the human ratings, we z-normalize the scores per rater and average them across segments. We use bootstrap resampling (Koehn, 2004) to compute 95% confidence

[2] `https://hf.co/Unbabel/XCOMET-XL`

[3] For Vallader fluency and Sursilvan accuracy, we initially observed low IAA and, after consulting with the respective raters, excluded three annotation runs from the final dataset and collected new ratings for Sursilvan accuracy from a third and fourth individual. Therefore, the final number of raters included in the results is 17.

intervals.

6 Results

Tables 4 and 5 report the results of the automatic evaluation for translation directions. Figure 4 visualizes the results of the human evaluation of the top three systems; detailed per-variety results are reported in Appendix M.

HR→LR augmentation does not benefit the lowest-resource varieties. The benefit of HR→LR augmentation can be measured by comparing a model that was only trained on natively parallel data ("No data augmentation") with a model that was trained on both the native and the synthetic parallel data. Even though the HR→LR augmented model is trained on 14× more data (Appendix H), it does not consistently outperform the baseline. In the German→Romansh direction (Table 4), the model does improve on the higher-resource varieties (Rumantsch Grischun, Sursilvan and Vallader), but is worse in the lower-resource varieties (Sutsilvan, Surmiran and Puter), with a decline of -17 BLEU in the case of Sutsilvan. This suggests that the synthetic text generated through HR→LR augmentation is not helpful for fine-tuning an MT system on these lower-resource varieties. An analysis of the synthetic text in Section 7 will corroborate this hypothesis.

HR diversity does not help HR→LR. As expected, HR→LR augmentation is beneficial in the Romansh→German direction (Table 5), with an average improvement over the no-augmentation baseline of 7.9–9.4 COMET. Comparing the effect of the two source datasets (Europarl and Newscrawl/FineWeb2), we find that the latter dataset gives a consistent improvement in the Romansh→German direction (+1.5 COMET on average), but has no consistent effect in the German→Romansh direction (-0.2 BLEU). Thus, the domain diversity of the source text appears to be a minor factor for HR→LR augmentation.

LR→HR augmentation is superior for both downstream MT directions. An MT model fine-tuned on LR→HR synthetic data outperforms the models trained on HR→LR synthetic data and those without any data augmentation. This is true for both test-time translation directions and for all varieties of Romansh, with average gains of 15.8 BLEU for German→Romansh and 1.1 COMET for Romansh→German.

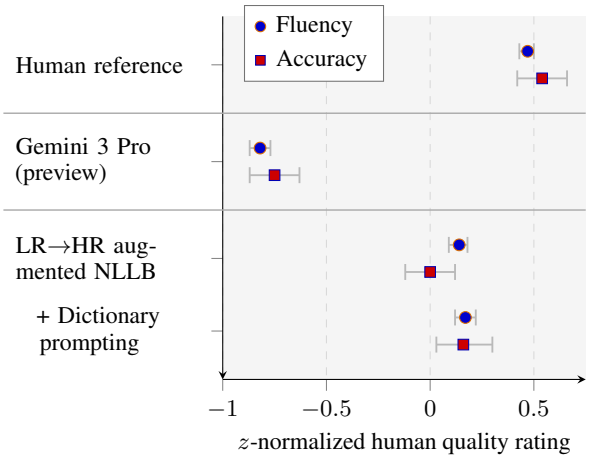

Figure 4: Human ratings for German→Romansh translation, averaged across the six target varieties. The error bars show 95% bootstrap confidence intervals.

A comparison of COMET scores and BLEU scores for Romansh→German (Appendix I) indicates that the bidirectionality of the improvement could be explained by an increased semantic accuracy of the parallel data: On average, HR→LR-augmented models have a similar German BLEU score as LR→HR-augmented models, but they are clearly worse in terms of COMET. Since COMET is more likely to pick up long-tail semantic errors than BLEU, this phenomenon could explain the disagreement between the metrics.

Overall, these results suggest that LLM-based data augmentation should ideally be performed in the LR→HR direction, based on authentic monolingual data in the low-resource language, and that work to collect such data will continue to contribute to MT quality.

LLM-based data augmentation enables variety-aware MT into Romansh. Our fine-tuned NLLB model outperforms Gemini 3 Pro by 9.9 BLEU on average for German→Romansh translation, with gains of up to 23.2 BLEU for the lowest-resource variety, Sutsilvan (Table 4). The human evaluation (Figure 4) shows that speakers rate our model substantially higher than Gemini in fluency. At the system level, they also prefer our model for accuracy, although item-level agreement on accuracy is only moderate (Table 3). In the Romansh→German direction, our model still underperforms Gemini in terms of COMET, which

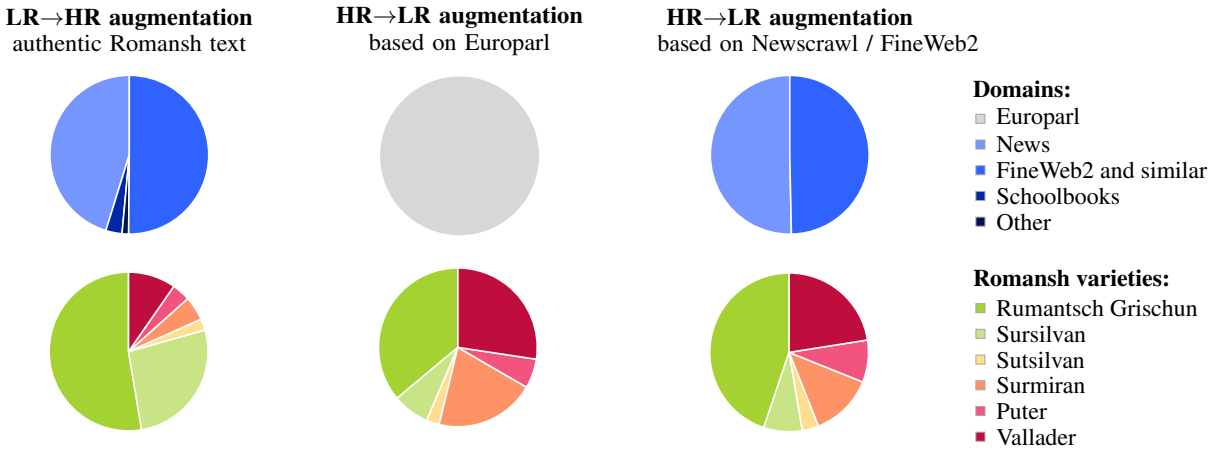

Figure 5: Comparison of Romansh data (authentic / synthetic) with respect to the included domains (top row) and the proportion of Romansh varieties according to LID classification (bottom row).

Dataset	LID agreement							Dictionary coverage						
	RG	Surs.	Suts.	Surm.	Puter	Vall.	Avg.	RG	Surs.	Suts.	Surm.	Puter	Vall.	Avg.
Authentic Romansh text	-	-	-	-	-	-	-	90.5	87.9	88.5	87.7	83.9	87.1	87.6
HR→LR synthetic datasets:														
– based on Europarl	100.0	40.3	17.3	80.9	33.1	95.8	61.2	93.5	80.3	58.9	71.5	77.9	83.0	77.5
– Newscrawl / FineWeb2	99.2	41.5	17.0	61.6	47.0	86.9	58.9	89.2	74.9	57.5	66.0	72.9	76.6	72.9

Table 6: Computational analysis of Romansh data (authentic / synthetic) with regard to adherence to variety-specific features. **LID agreement** quantifies the proportion of synthetic samples where the requested target variety matches the prediction of an LID system. **Dictionary coverage** is the proportion of words in the text that overlap with variety-specific Romansh dictionaries.

underlines the asymmetry of Gemini's capabilities.

Including dictionary information in the data augmentation prompt did slightly improve the downstream model both in the automatic evaluation (+0.1 BLEU / +0.1 COMET on average) and the human evaluation, but below statistical significance and without consistency across the individual varieties. Future work could employ the available dictionary information beyond prompting, for instance for rejection fine-tuning.

The low performance of HR→LR augmentation for German→Romansh translation raises the question what specific deficiencies there are in the synthesized Romansh text and how it compares to the authentic text that is used for LR→HR augmentation. We perform a computational analysis using a language identification (LID) system (Model et al., 2026) and a dictionary look-up software that matches word forms in Romansh text with variety-specific dictionaries (Fischer et al., 2026).

7 Analysis of Augmented Data

LLMs struggle to synthesize identifiable text in low-resource varieties. The left-hand side of Table 6 shows that only 17% of the synthetic Sutsilvan samples are identified as Sutsilvan by the LID system, and less than 50% of synthetic Puter samples are identified as Puter. This is consistent with the confusion matrix of Gemini 2.5 Flash in Figure 3, where translations into Sutsilvan and Puter score higher against RG and Vallader references than the references for the actual target varieties.

However, a potential confound of such an LID analysis is that LID might be inherently easier for some varieties. For example, the relatively high LID agreement of synthetic Surmiran (61.6–80.9%) could be explained by the especially high recall of the LID classifier on Surmiran (Model et al., 2026).

When training the MT system, we make sure to label the samples using the LID system, overriding the LLM-produced label. As a consequence, the HR→LR augmented data are imbalanced with regard to the varieties (Figure 5), albeit to a lesser degree than the available authentic Romansh data.

LLMs struggle to generate variety-specific word forms. The right-hand side of Table 6 indicates that the synthetic Romansh text contains fewer

words that match with a dictionary than the authentic Romansh text. For example, only 57.5–58.9% of words in synthetic Sutsilvan text can be found in a Sutsilvan dictionary, whereas 88.5% of words in authentic Sutsilvan text can be identified.

Taken together, the analysis indicates that HR→LR augmented data are less useful for MT fine-tuning because they reflect the LLM's bias towards higher-resource varieties and limited exposure to the lower-resource varieties.

8 Conclusion

Our case study on six Romansh language varieties shows that translation direction matters when an LLM is used for data augmentation. Back-translation from the lower-resource language produces better training signals than forward translation from the higher-resource language. This suggests that collecting authentic monolingual data may remain essential in similar low-resource settings.

Limitations

We limit our controlled data augmentation experiments to Gemini 2.5 Flash and greedy decoding. Gemini 2.5 Pro performed slightly better than Flash in preliminary experiments, but using it for data augmentation with all available data would have exceeded our budget. Experiments with larger models and alternative sampling strategies thus remain future work. Additionally, we did not test dictionary prompting with forward translation from German, due to the high cost of LLM usage for brute-force dictionary prompting (Appendix G).

We trained five models, two of which were included in the human evaluation. Alternative data augmentation strategies were evaluated intrinsically by translating samples from the WMT24++ benchmark and computing COMET (Table 1), without training and evaluating a downstream MT model.

Our evaluation focuses on German→Romansh and Romansh→German translation. Future work could also explore translation *between* Romansh varieties, as well as data augmentation for these directions. The human evaluation study was limited to the German→Romansh direction, since this is the most challenging direction and the direction for which no trained MT metrics are currently available. Finally, a limitation of the study is that it does not analyze document-level errors in a targeted way, such as inconsistency in terminology and named entities. However, the effect size in Figure 4 makes it unlikely that document-level aspects would affect the overall ranking between the systems we compared.

Ethical considerations

We release our trained model under a CC-BY-NC 4.0 license, in line with the original license of NLLB. For training, we use a mix of publicly available data (links in Appendix F), and a number of datasets that were made available to us for research use (listed in Appendix F without link).

Regarding privacy, the web-crawled portion of our training data (FineWeb2 and FinePDFs) has email addresses and IP addresses anonymized by default. The other datasets we used, such as news articles, schoolbooks, and dictionaries, are at low risk of containing personally identifiable information.

For the human evaluation, raters were personally recruited based on their native Romansh proficiency. They were compensated at a standard hourly rate. We release the human evaluation data without any personally identifiable information.

Acknowledgments

We thank RTR and Fundaziun Patrimoni Cultural RTR for their support. We are grateful to Zachary Hopton, Diana Merkle, Anna Rutkiewicz and Sudehsna Sivakumar for help with data curation, Uniun dals Grischs for contributing dictionary data for Puter and Vallader, and Giuanna Caviezel, Not Soliva and their seminar participants for helpful feedback. We also acknowledge the contribution of the native speakers who participated in the human evaluation study, namely Claudia Cadruvi, Martina Caprez, Marietta Cathomas, Eliane Cathomen, Laura Decurtins, Andri Florineth, Arina Lazzarini, Viviana Lazzarini, Lea Livers, Patrick Meister, Gierina Michael, Bettina Nicca, Zegna Pittet-Dosch, Barbara Riesch, Manuela Schnoz-Flury, and Annalea Stuppan, and those who would rather not be named. For this publication, use was made of media data made available via SwissdoxLiRI by the Linguistic Research Infrastructure of the University of Zurich (see https://www.liri.uzh.ch/en/services/swissdox.html for more information).

References

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2020. Domain, translationese and noise in synthetic data for neural machine translation. *Preprint*, arXiv:1911.03362.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Eva Caviezel. 1993. Geschichte von Verschriftung, Normierung und Standardisierung des Surselvischen. *Societad Retorumantscha*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Renata Coray. 2008. *Von der Mumma Romontscha zum Retortenbaby Rumantsch Grischun : rätoromanische Sprachmythen*. Cultura alpina. Institut für Kulturforschung Graubünden, Chur.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Sara Court and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA. Association for Computational Linguistics.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource MT via synthetic data generation with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27674–27692, Suzhou, China. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Dominic P. Fischer, Zachary Hopton, and Jannis Vamvas. 2026. RUMLEM: A dictionary-based lemmatizer for Romansh. In *Proceedings of the 11th Edition of the Swiss Text Analytics Conference*, pages 125–132, Zurich, Switzerland. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Frontull and Georg Moser. 2024. Rule-based, neural and LLM back-translation: Comparative insights from a variant of Ladin. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 128–138, Bangkok, Thailand. Association for Computational Linguistics.

Jean-Jacques Furer. 2005. *Die aktuelle Lage des Romanischen*. Office fédéral de la statistique.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Matthias Grünert. 2024. Rätoromanisch. In Elvira Glaser, Johannes Kabatek, and Barbara Sonnenhauser, editors, *Sprachenräume der Schweiz. Band 1: Sprachen*, pages 156–184. Narr Francke Attempto, Tübingen.

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

Zachary William Hopton, Jannis Vamvas, Andrin Büchler, Anna Rutkiewicz, Rico Cathomas, and Rico Sennrich. 2026. The mediomatix corpus: Parallel data for Romansh language varieties via comparable schoolbooks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 290–306, Rabat, Morocco. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 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, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. 2025. Finepdfs. `https://huggingface.co/datasets/HuggingFaceFW/finepdfs`.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340, Vienna, Austria. Association for Computational Linguistics.

Ricarda Liver. 2010. *Rätoromanisch: eine Einführung in das Bündnerromanische*, 2., überarbeitete und erweiterte Auflage edition. Narr Studienbücher. Narr Verlag, Tübingen.

Charlotte Model, Sina Ahmadi, and Jannis Vamvas. 2026. Robust language identification for Romansh varieties. In *Proceedings of the 11th Edition of the Swiss Text Analytics Conference*, pages 101–110, Zurich, Switzerland. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. SwiLTra-bench: The Swiss legal translation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.

Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh

Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, and 12 others. 2026. Omnilingual MT: Machine translation for 1,600 languages. *Preprint*, arXiv:2603.16309.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo. 2025. Finewiki. Source: Wikimedia Enterprise Snapshot API (https://api.enterprise.wikimedia.com/v2/snapshots). Text licensed under CC BY-SA 4.0 with attribution to Wikipedia contributors.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pásztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan Garcia Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, and 84 others. 2026. Apertus: Democratizing open and compliant LLMs for global language environments. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46877–46955, San Diego, California, United States. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz, and Rico Sennrich. 2025. Expanding the WMT24++ benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1028–1047, Suzhou, China. Association for Computational Linguistics.

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 the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A NLLB Fine-tuning Hyperparameters

Hyperparameter	Value
Base model	NLLB-200-Distilled 1.3B
Learning rate	2×10^{-4}
Warmup updates	150
Max epochs	10
Stopping criterion	Validation loss on a held-out split of the training data
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$)
LR scheduler	Inverse square root
Dropout	0.1
Label smoothing	0.1
Gradient clipping	1.0
Weight decay	0.01
Mixed precision	bfloat16
Max tokens per batch	4096
Update frequency	8
Effective batch size	32,768 tokens
Max sequence length	512 tokens
Hardware	Single H100 GPU (80GB memory)

Table 7: Fine-tuning hyperparameters.

B NLLB Decoding

For NMT inference, we use the CTranslate2 toolkit[4] with beam search (beam size 4) and a maximum decoding length of 256 tokens. Input texts are segmented before translation: For German→Romansh, we split German into sentences using spaCy with the `de_dep_news_trf` model; for Romansh→German, we split Romansh into text segments using SaT (Frohmann et al., 2024), with model `sat-12l-sm` and a maximum segment length of 500 characters.

C Data Preprocessing

Monolingual Data Monolingual data were preprocessed on the document level before performing the LLM-based data augmentation. We remove documents/segments with excessive punctuation (>50% punctuation tokens), collapse repeated punctuation (max 3 consecutive), filter short documents (<5 tokens), split long documents (>500 tokens) along newline boundaries, and deduplicate.

Parallel Data Parallel data (both synthetic and natively parallel) underwent similar filtering: we remove pairs with excessive punctuation, collapse repeated punctuation, filter non-Romansh text using a dictionary-based heuristic (at least 50% of the words should be found in a Romansh dictionary), and deduplicate. We then segment the documents into lines based on newline boundaries, align them across languages using Vecalign (Thompson and Koehn, 2019) with embeddings from OpenAI's `text-embedding-3-small` model, and filter out pairs with a length ratio greater than 1.5.

Romansh Variety Classification We classify all Romansh text (including synthetic text generated through forward translation, but excluding the parallel data derived from dictionaries) into one of 6 Romansh varieties using a Support Vector Machine classifier released by Model et al. (2026).

D Details on Dictionary Prompting

We use Romansh–German dictionaries for all six Romansh varieties. For Vallader and Puter, we use dictionaries provided by Uniun dals Grischs.[5] For the other varieties, we use dictionaries from the Pledari Grond project.[6]

For each word in the source text, we include the following information in the prompt: the Romansh word form as it appears in the text, the inferred lemma, all corresponding German lemmas, and all possible morphological analyses of the word form based on paradigms provided by the dictionaries. We used the `rumlem` software (Fischer et al., 2026) to derive this information from the Romansh text. We do not use context-sensitive lemmatization, which we defer to future work. Similarly, we do not inflect the German lemmas.

For consistency, the few-shot examples in the dictionary prompting setup also include dictionary entries.

To reduce prompt length, we exclude dictionary information for high-frequency words, specifically those with a rank below 500 per variety in our monolingual datasets.

We also experimented with approaches requiring less dictionary information, such as providing relevant dictionary entries only in a follow-up prompt based on an unconstrained translation created without dictionary information (cf. Bogoychev and Chen, 2023). However, preliminary experiments indicated that simply providing all entries in the initial prompt works best in terms of BLEU and COMET on our validation set.

E Human Evaluation Methodology

We use a customized version of the Appraise platform (Federmann, 2018). Raters are shown two

[4] https://github.com/OpenNMT/CTranslate2

[5] https://www.udg.ch/dicziunari

[6] https://pledarigrond.ch

system outputs per segment (for fluency) or per document (for accuracy). Fluency is evaluated monolingually, i.e., raters do not see the source text, while accuracy is evaluated bilingually. The systems are selected randomly per segment (for fluency) or per document (for accuracy) out of a set of four: the human reference from the WMT24++ benchmark (Vamvas et al., 2025), Gemini 3 Pro, NLLB trained with the baseline LR→HR augmentation approach, and NLLB trained with LR→HR augmentation using dictionary prompting. Screenshots of the evaluation interface are shown in Appendices O.1 and O.2; the evaluation guidelines are provided in Appendix O.3.

In addition to the document-level accuracy ratings, we encouraged raters to indicate segment-level pairwise preferences, by selecting the better translation among the two systems. The main human evaluation results are reported in Figure 4; segment-level pairwise win-rates are reported in Appendix M.

F Datasets used for MT Training

F.1 Parallel Datasets

Name	Romansh varieties	URL	RM tokens	DE tokens
Dictionary data	RG, Surs., Suts., Surm., Puter, Vall.	(↗)	3,298,051	6,530,231
Mediomatix (Hopton et al., 2026)	Surs., Suts., Surm., Puter, Vall.	↗	† 6,946,947	-
Press releases of Canton Grisons	RG	(↗)	1,840,537	1,397,196
SwissLawTranslations (Niklaus et al., 2025)	RG	↗	607,517	494,980
Laws of Canton Grisons	RG	↗	538,174	402,047
Parallel data contributed by RTR	RG, Surs., Suts., Surm., Puter, Vall.	-	646,856	-
Storyweaver	Surs., Suts., Surm., Puter, Vall.	-	48,475	9,170
Total			13,926,557	8,833,624

Table 8: Parallel datasets included in the MT training. Tokens are counted based on whitespace tokenization. † Counting each pairwise combination of varieties separately.

F.2 Monolingual Datasets

Name	URL	Tokens
Europarl (Koehn, 2005)	↗	(deduplicated) 52,239,860
Newscrawl 2023 (Kocmi et al., 2025)	↗	∼894,960,000
FineWeb2 (Penedo et al., 2025)	↗	∼262,271,052,000

Table 9: Monolingual German datasets used for the two **HR→LR augmentation** variants. Token counts are approximate. Note that we sample a subset from these datasets, independently for each Romansh target variety.

Name	Romansh varieties	URL	Tokens
FineWeb2 (Penedo et al., 2025)	RG	↗	48,340,896
La Quotidiana (1997–2008, 2021–2025)	RG, Surs., Suts., Surm., Puter, Vall.	↗	38,993,608
News articles by RTR	RG	-	22,705,038
FinePDFs (Kydlíček et al., 2025)	RG	↗	18,856,327
Mediomatix (unaligned) (Hopton et al., 2026)	Surs., Suts., Surm., Puter, Vall.	↗	4,639,563
FineWiki (Penedo, 2025)	RG	↗	2,827,954
Audio transcriptions by RTR	RG, Surs., Suts., Surm., Puter, Vall.	(↗)	1,467,957
Theater plays	RG, Surs., Suts., Surm., Puter, Vall.	↗	1,079,943
Municipal documents	Surs., Suts., Surm., Vall.	↗	318,308
Historical Dictionary of Switzerland	RG	(↗)	234,715
Revista digl noss Sulom	Surm.	-	187,247
Babulins	RG, Surs., Suts., Surm., Puter, Vall.	-	49,444
Total			139,701,000

Table 10: Monolingual Romansh datasets used for **LR→HR augmentation**. Tokens are counted based on whitespace tokenization.

G Statistics of LLM Requests

	Requests	Total input tokens	Total output tokens	Approx. API cost
HR→LR augmentation				
– Europarl	251,432	283,562,126	252,592,957	350 USD
– Newscrawl / FineWeb2	2,615,424	1,163,299,838	267,593,106	500 USD
LR→HR augmentation				
– Baseline prompt	765,908	530,210,202	194,726,402	315 USD
– Dictionary prompting	765,908	8,292,343,080	196,047,179	1260 USD

Table 11: Statistics of LLM requests for different data augmentation strategies. Tokens are counted in terms of the tokenizer of Gemini 2.5 Flash. The variants required a different number of requests (and consequently, few-shot tokens) because the average document length can vary. Furthermore, different tokenizer fertilities for German and Romansh lead to different input/output token ratios.

H Statistics of Preprocessed Training Data

The following tables report training data statistics per language, before upsampling or reversing the translation directions for bidirectional training. Percentages in parentheses indicate the proportion of segments/tokens where either the sample itself or the corresponding source or target segment is synthetic. Tokens are counted based on whitespace tokenization.

Language	Europarl		Newscrawl / FineWeb2	
	# Segments	# Tokens	# Segments	# Tokens
German	7.8M (70% synthetic)	121.0M (94% synthetic)	6.5M (64% synthetic)	118.7M (94% synthetic)
RG	2.7M (71% synthetic)	49.8M (92% synthetic)	2.7M (70% synthetic)	62.7M (94% synthetic)
Sursilvan	0.7M (60% synthetic)	11.1M (85% synthetic)	0.6M (55% synthetic)	12.0M (86% synthetic)
Sutsilvan	0.6M (25% synthetic)	5.0M (66% synthetic)	0.6M (24% synthetic)	6.1M (72% synthetic)
Surmiran	1.4M (83% synthetic)	26.8M (97% synthetic)	0.8M (70% synthetic)	17.8M (96% synthetic)
Puter	0.8M (40% synthetic)	9.3M (81% synthetic)	0.8M (42% synthetic)	13.0M (86% synthetic)
Vallader	2.0M (74% synthetic)	36.5M (95% synthetic)	1.4M (63% synthetic)	31.2M (94% synthetic)
Total	16.1M (68% synthetic)	259.5M (93% synthetic)	13.4M (62% synthetic)	261.4M (93% synthetic)

Table 12: Training data statistics for the HR→LR direction.

Language	Baseline prompt		Dictionary prompting	
	# Segments	# Tokens	# Segments	# Tokens
German	6.3M (63% synthetic)	115.3M (94% synthetic)	6.4M (64% synthetic)	117.5M (94% synthetic)
RG	3.3M (76% synthetic)	67.1M (94% synthetic)	3.4M (76% synthetic)	68.9M (94% synthetic)
Sursilvan	1.1M (74% synthetic)	33.1M (95% synthetic)	1.1M (75% synthetic)	33.9M (95% synthetic)
Sutsilvan	0.5M (14% synthetic)	4.4M (61% synthetic)	0.5M (14% synthetic)	4.5M (61% synthetic)
Surmiran	0.4M (37% synthetic)	6.7M (89% synthetic)	0.4M (38% synthetic)	6.8M (89% synthetic)
Puter	0.7M (25% synthetic)	6.2M (71% synthetic)	0.7M (25% synthetic)	6.3M (71% synthetic)
Vallader	0.8M (34% synthetic)	13.3M (86% synthetic)	0.8M (35% synthetic)	13.5M (86% synthetic)
Total	13.1M (61% synthetic)	246.1M (92% synthetic)	13.3M (62% synthetic)	251.3M (93% synthetic)

Table 13: Training data statistics for the LR→HR direction.

I Automatic Evaluation: Romansh→German BLEU Scores

System	WMT24++						BOUQuET	Avg.
	RG	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader	RG	
Gemini 2.5 Flash	55.7	50.0	44.9	49.9	52.0	58.9	**33.9**	49.3
Gemini 3 Pro (preview)	**55.8**	**50.8**	**48.7**	52.2	**53.8**	**60.8**	33.8	**50.8**
Gemini 3.7 Flash	55.2	50.3	48.2	**52.9**	53.3	60.6	33.6	50.6
Fine-tuned NLLB								
No data augmentation	38.3	33.9	33.3	33.7	34.6	37.2	26.5	33.9
HR→LR augmentation:								
– Europarl data	45.8	42.2	40.9	43.9	45.5	49.9	30.1	42.6
– Newscrawl / FineWeb2	50.5	46.1	44.5	47.4	49.6	54.7	33.1	46.6
LR→HR augmentation:								
– Baseline prompt	50.1	46.0	43.9	46.8	49.3	54.8	33.0	46.3
– Dictionary prompting	50.5	46.3	44.2	47.3	49.5	54.8	33.3	46.5

Table 14: Per-variety BLEU scores for Romansh→German translation on the second half of the WMT24++ benchmark, together with BOUQuET results for Rumantsch Grischun. **Bold** denotes the best system; underline indicates the best variant among the fine-tuned NLLB models; the last column reports the arithmetic mean of the seven columns.

J Prompt Templates

J.1 Baseline prompt

```
Translate the following text into {target_language}. The text is written in {source_language} and is provided to you wrapped in triple backticks (```). Just answer with the translation and nothing else.\n\n```{source_document}```\n
```

J.2 Prompt with Dictionary Information

```
Translate the text below into {target_language}. The text is written in {source_language} and is provided to you wrapped in triple backticks (```).\n\nSome of the dictionary entries below might be helpful for translating the text:\n\n{dictionary_entries}\n\nJust answer with the translation and nothing else.\n\nText to translate:\n```{source_segment}```
```

K Full Example for a Prompt with Dictionary Information

Translate the text below into German. The text is written in Romansh (Sursilvan variety) and is provided to you wrapped in triple backticks (```).

Some of the dictionary entries below might be helpful for translating the text:

- stos (lemma: stuer) → müssen [PoS=V;Mood=IND;Tense=PRS;Person=2;Number=SG]
- violenza (lemma: violenza) → Heftigkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Gewalt [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Stärke [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Wucht [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Gewalttätigkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Tätlichkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Vergewaltigung [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Notzucht [PoS=N;Gender=FEM;Number=SG]
- meina (lemma: meina) → führen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → ausführen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → transportieren [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → leiten [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → anführen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → vorstehen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → regieren [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → verbringen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → durchführen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → vollenden [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → verwalten [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → fahren [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → steuern [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → lenken [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → wenden [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → drehen [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → kehren [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → richten [PoS=V;Finiteness=NFIN]
- meina (lemma: meina) → decken [PoS=V;Finiteness=NFIN]
- violenza (lemma: violenza) → Heftigkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Gewalt [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Stärke [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Wucht [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Gewalttätigkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Tätlichkeit [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Vergewaltigung [PoS=N;Gender=FEM;Number=SG]
- violenza (lemma: violenza) → Notzucht [PoS=N;Gender=FEM;Number=SG]
- decisiun (lemma: decisiun) → Entscheid [PoS=N;Gender=FEM;Number=SG]
- decisiun (lemma: decisiun) → Entscheidung [PoS=N;Gender=FEM;Number=SG]
- decisiun (lemma: decisiun) → Abmachung [PoS=N;Gender=FEM;Number=SG]
- decisiun (lemma: decisiun) → Entschluss [PoS=N;Gender=FEM;Number=SG]
- decisiun (lemma: decisiun) → Entschliessung [PoS=N;Gender=FEM;Number=SG]
- dedicau (lemma: dedicar) → widmen [PoS=V;VerbForm=PTCP;Tense=PST;Gender=NEUT]
- protecziun (lemma: protecziun) → Schutz [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Beschützung [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Bewahrung [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Wehr [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Protektion [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Begünstigung [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Förderung [PoS=N;Gender=FEM;Number=SG]
- protecziun (lemma: protecziun) → Gönnerschaft [PoS=N;Gender=FEM;Number=SG]

Just answer with the translation and nothing else.

Text to translate:
```Quei hai jeu buca detg. Quei stos ti era buc. Ti sas meglier che tut ils auters che violenza meina mintga ga a dapli violenza. Il cussegl ha priu la decisiun. Il cussegl dat ei buca pli! Nus havein dedicau noss’entira veta, mellis onns, alla protecziun dalla veta.```
```

L Generative Confusion Matrices

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	40.3	18.4	8.9	11.8	12.2	15.4
Surs.	27.2	28.0	8.6	10.5	10.0	12.7
Suts.	22.0	16.4	10.6	12.6	10.0	12.1
Surm.	21.6	13.7	9.4	16.9	11.2	12.5
Puter	14.8	10.6	7.5	8.6	22.0	24.6
Vall.	16.3	11.1	7.6	8.7	20.6	27.3

(a) Gemini 2.5 Flash

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	45.3	19.8	9.1	12.8	12.6	16.2
Surs.	21.7	37.1	9.0	10.5	9.2	11.1
Suts.	16.4	14.3	17.3	14.9	10.0	11.1
Surm.	16.5	12.5	10.2	27.1	10.0	10.7
Puter	13.6	9.8	7.1	9.2	34.7	24.1
Vall.	17.3	11.5	7.9	9.7	23.2	36.1

(b) Gemini 3 Pro

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	40.9	18.1	8.6	11.5	12.0	15.0
Surs.	22.5	34.2	8.8	10.4	9.5	11.6
Suts.	16.3	14.5	14.8	13.5	9.6	10.9
Surm.	16.7	12.0	10.0	22.1	10.7	11.2
Puter	13.9	10.3	7.4	8.8	30.4	22.1
Vall.	15.3	10.7	7.3	8.5	19.8	29.8

(c) Gemini 3.7 Flash

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	35.9	16.5	7.4	10.4	10.8	13.3
Surs.	23.1	29.4	7.4	8.6	8.6	10.1
Suts.	11.1	9.0	29.2	8.9	6.5	7.4
Surm.	16.6	10.8	8.6	25.3	8.2	9.1
Puter	12.9	8.8	6.2	7.4	31.0	19.2
Vall.	19.7	11.6	6.6	8.2	16.9	26.0

(d) No data augmentation

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	40.8	17.3	7.8	11.1	11.4	14.7
Surs.	23.6	31.3	7.7	9.4	8.7	11.1
Suts.	21.6	14.8	12.4	13.0	9.8	12.0
Surm.	22.0	12.7	9.4	18.7	10.9	12.2
Puter	16.4	10.5	6.7	8.2	23.8	27.4
Vall.	17.0	10.6	6.8	8.2	20.9	30.7

(e) HR→LR augmentation with Europarl data

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	41.4	17.9	8.2	11.2	11.6	15.1
Surs.	24.0	30.4	7.9	9.8	9.3	11.4
Suts.	19.6	15.5	12.1	13.2	9.7	11.7
Surm.	20.5	13.2	9.2	18.5	10.6	11.7
Puter	15.1	10.1	6.9	8.4	24.6	25.4
Vall.	16.5	10.6	7.0	8.2	21.0	29.4

(f) HR→LR augmentation with Newscrawl / FineWeb2 data

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	48.1	19.9	8.5	12.2	12.4	15.7
Surs.	18.9	44.3	8.3	9.4	8.2	9.9
Suts.	8.9	9.2	40.6	9.9	6.7	7.4
Surm.	12.7	10.5	9.7	42.3	8.2	8.9
Puter	11.8	8.8	6.3	8.0	44.8	24.8
Vall.	14.6	9.9	6.7	8.0	25.0	44.3

(g) LR→HR augmentation

↓ tgt	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	48.4	19.9	8.6	12.2	12.4	15.5
Surs.	19.0	44.5	8.3	9.4	8.2	9.8
Suts.	9.0	9.3	40.5	10.1	6.7	7.5
Surm.	12.5	10.5	9.7	43.0	8.1	8.9
Puter	11.9	8.9	6.3	8.0	44.9	25.0
Vall.	14.8	10.1	6.7	8.1	24.7	44.6

(h) LR→HR augmentation with Dictionary prompting

Figure 6: Confusion matrices similar to Figure 3 illustrating the target variety adherence in German→Romansh translation. Results are based on BLEU.

M Detailed Human Evaluation Results

M.1 Fluency

System	RG	Surs.	Suts.	Surm.	Puter	Vall.
Human reference	0.24 ± 0.10	0.66 ± 0.07	0.62 ± 0.07	0.60 ± 0.08	0.50 ± 0.07	0.23 ± 0.12
Gemini 3 Pro (preview)	-0.30 ± 0.14	-0.75 ± 0.11	-1.35 ± 0.08	-0.98 ± 0.11	-0.92 ± 0.15	-0.65 ± 0.13
LR→HR augmented NLLB	0.03 ± 0.11	0.02 ± 0.12	0.36 ± 0.09	0.17 ± 0.10	0.05 ± 0.10	0.19 ± 0.12
+ Dictionary prompting	0.04 ± 0.11	0.10 ± 0.12	0.26 ± 0.09	0.28 ± 0.11	0.15 ± 0.12	0.20 ± 0.14

Table 15: Per-variety fluency scores for German→Romansh translation. Ratings were made on the segment level and then averaged across segments.

M.2 Accuracy

System	RG	Surs.	Suts.	Surm.	Puter	Vall.
Human reference	0.40 ± 0.27	0.74 ± 0.25	0.62 ± 0.26	0.64 ± 0.22	0.31 ± 0.31	0.55 ± 0.21
Gemini 3 Pro (preview)	-0.25 ± 0.34	-0.85 ± 0.17	-0.75 ± 0.18	-1.16 ± 0.31	-0.92 ± 0.23	-0.57 ± 0.33
LR→HR augmented NLLB	-0.02 ± 0.24	-0.08 ± 0.27	-0.16 ± 0.27	-0.01 ± 0.27	0.40 ± 0.26	-0.15 ± 0.30
+ Dictionary prompting	-0.03 ± 0.33	0.21 ± 0.24	0.36 ± 0.29	0.03 ± 0.26	0.18 ± 0.28	0.19 ± 0.27

Table 16: Per-variety accuracy scores for German→Romansh translation. Ratings were made on the document level and then averaged across documents.

M.3 Pairwise Win-Rates of Systems in Human Evaluation

	Human reference	Gemini 3 Pro (preview)	LR→HR augmented NLLB	+ Dictionary prompting
Human reference	-	72 ± 4	53 ± 4	56 ± 5
Gemini 3 Pro (preview)	9 ± 2	-	15 ± 3	11 ± 3
LR→HR augmented NLLB	23 ± 3	63 ± 4	-	20 ± 4
+ Dictionary prompting	22 ± 4	66 ± 4	22 ± 4	-

Table 17: Win-rates of systems in segment-level fluency evaluation, on average across varieties. The values indicate the percentage of segments where the system in the row was rated higher than the system in the column.

	Human reference	Gemini 3 Pro (preview)	LR→HR augmented NLLB	+ Dictionary prompting
Human reference	-	80 ± 10	66 ± 10	50 ± 14
Gemini 3 Pro (preview)	5 ± 4	-	4 ± 4	11 ± 5
LR→HR augmented NLLB	11 ± 7	64 ± 10	-	26 ± 9
+ Dictionary prompting	15 ± 8	76 ± 12	25 ± 9	-

Table 18: Win-rates of systems in document-level accuracy ranking, on average across varieties.

	Human reference	Gemini 3 Pro (preview)	LR→HR augmented NLLB	+ Dictionary prompting
Human reference	-	60 ± 5	51 ± 4	51 ± 7
Gemini 3 Pro (preview)	10 ± 3	-	14 ± 3	14 ± 3
LR→HR augmented NLLB	20 ± 3	54 ± 4	-	19 ± 3
+ Dictionary prompting	24 ± 6	56 ± 4	22 ± 3	-

Table 19: Win-rates of systems in segment-level accuracy evaluation, on average across varieties.

N Human Evaluation Statistics

N.1 Number of Ratings

Variety	Human reference	Gemini 3 Pro (preview)	LR→HR augmentation	+ Dictionary prompting
Rumantsch Grischun	**266 / 38 / 92**	**280 / 54 / 149**	**292 / 40 / 166**	**279 / 36 / 124**
Rater 1	243 / 33 / 59	249 / 42 / 96	268 / 30 / 114	258 / 31 / 102
Rater 2	23 / 5 / 33	31 / 12 / 53	24 / 10 / 52	21 / 5 / 22
Sursilvan	**250 / 46 / 188**	**260 / 49 / 176**	**245 / 54 / 179**	**260 / 55 / 140**
Rater 3	208 / 0 / 0	210 / 0 / 0	190 / 0 / 0	204 / 0 / 0
Rater 4	0 / 37 / 112	0 / 40 / 96	0 / 42 / 118	0 / 43 / 87
Rater 5	0 / 9 / 76	0 / 9 / 80	0 / 12 / 62	0 / 12 / 54
Rater 6	27 / 0 / 0	31 / 0 / 0	33 / 0 / 0	36 / 0 / 0
Rater 7	15 / 0 / 0	19 / 0 / 0	22 / 0 / 0	20 / 0 / 0
Sutsilvan	**250 / 36 / 180**	**231 / 40 / 184**	**232 / 47 / 196**	**274 / 43 / 144**
Rater 8	69 / 14 / 86	71 / 18 / 91	63 / 21 / 94	73 / 21 / 70
Rater 9	54 / 22 / 90	27 / 22 / 93	48 / 26 / 96	57 / 22 / 70
Rater 10	127 / 0 / 2	133 / 0 / 0	121 / 0 / 6	144 / 0 / 4
Surmiran	**245 / 51 / 246**	**263 / 29 / 98**	**280 / 48 / 118**	**229 / 40 / 206**
Rater 11	132 / 36 / 134	139 / 22 / 58	147 / 36 / 66	123 / 28 / 107
Rater 12	113 / 15 / 112	124 / 7 / 41	133 / 12 / 52	106 / 12 / 99
Puter	**262 / 32 / 64**	**192 / 38 / 140**	**261 / 43 / 76**	**214 / 45 / 146**
Rater 13	83 / 32 / 58	71 / 38 / 132	81 / 43 / 64	69 / 45 / 130
Rater 14	171 / 0 / 0	112 / 0 / 0	170 / 0 / 0	132 / 0 / 0
Rater 15	8 / 0 / 5	9 / 0 / 8	10 / 0 / 13	13 / 0 / 16
Vallader	**212 / 37 / 75**	**200 / 37 / 110**	**220 / 48 / 180**	**195 / 44 / 118**
Rater 16	212 / 0 / 0	200 / 0 / 0	220 / 0 / 0	195 / 0 / 0
Rater 17	0 / 37 / 75	0 / 37 / 110	0 / 48 / 180	0 / 44 / 118

Table 20: Statistics for human evaluation: number of fluency ratings of segments / number of accuracy ratings of documents / number of accuracy preferences of segment pairs. Pairwise win-rates based on the latter are reported in Appendix M.

N.2 Detailed Inter-Rater Agreement Analysis

Metric	RG	Surs.	Suts.	Surm.	Puter	Vall.
Number of overlapping segments	51	64	139	238	160	(0)
Item-level Pearson correlation	52.5	73.2	82.4	65.9	64.0	—
Pairwise comparison agreement rate	63.8	66.7	77.9	68.2	65.8	—
System-level Spearman correlation	80.0	100.0	100.0	100.0	80.0	—

Table 21: Inter-rater agreement metrics for segment-level fluency ratings.

Metric	RG	Surs.	Suts.	Surm.	Puter	Vall.
Number of overlapping documents	16	21	37	23	(0)	(0)
Item-level Pearson correlation	23.5	35.4	17.4	50.4	—	—
Pairwise comparison agreement rate	56.2	52.4	62.2	65.2	—	—
System-level Spearman correlation	80.0	40.0	100.0	80.0	—	—

Table 22: Inter-rater agreement metrics for document-level accuracy ratings.

Metric	RG	Surs.	Suts.	Surm.	Puter	Vall.
Number of overlapping segments	159	271	341	303	(0)	(0)
Item-level Pearson correlation	43.0	45.3	36.1	44.6	—	—
Pairwise comparison agreement rate	56.0	63.5	52.2	59.4	—	—
System-level Spearman correlation	80.0	100.0	100.0	80.0	—	—

Table 23: Inter-rater agreement metrics for segment-level accuracy ratings.

O Human Evaluation Interface and Guidelines

O.1 Annotation Interface – Fluency

O.2 Annotation Interface – Accuracy

O.3 Human Evaluation Guidelines

General Information on the Evaluation

For each idiom you evaluate, you will complete two passes: Fluency and Accuracy. These are explained in more detail below. For each pass, you will receive a separate account. This means you have one account for Fluency and one account for Accuracy for each idiom you evaluate. Click on the idiom you would like to evaluate first. Once you have started a pass, you must complete it before you can start another one. Since your two accounts exist independently of each other, it is theoretically possible not to finish the Fluency pass before starting Accuracy. However, please always finish the Fluency pass first before starting the Accuracy pass! If you evaluate multiple idioms, we recommend evaluating both Fluency and Accuracy for the first idiom before moving on to the next idiom.

On the evaluation page, you will now see two translations of an originally German text. Read them carefully. When evaluating Accuracy, also display the German source text. Differences between the two translations are highlighted in yellow. More detailed guidelines for each pass can be found below. If at any point during the evaluation process you can no longer see your task, simply copy this link: Link to platform

Once you click "Submit" for a task, you can no longer access it.

Fluency

Here, the focus is on evaluating idiomatic usage and grammatical correctness. Pay attention to whether the translation is correct within the Romansh idiom you are evaluating—idiomatic, grammatically correct, and overall natural-sounding. Below each segment you will find a slider from 0 to 6. Based on your own judgment, taking into account the approaches described below, estimate the score you would like to assign to each segment. In particular, consider whether the translation corresponds to the idiom being evaluated, and deduct points accordingly for any mixing with other idioms.

Example: You are evaluating a text in Vallader and notice that many formulations come from Puter. These are correct in Puter, but do not exist in Vallader. For the "Fluency" slider, choose a maximum of 2—depending on how good the remaining formulations are. If, in this example, the entire text is actually Puter, then choose 0 instead. If idioms are mixed within a segment, evaluate it exclusively within the framework of the idiom being assessed: what is Sutsilvan in a Sursilvan text should be counted as "incorrect." Invented terms or conjugations borrowed from other languages are also considered "incorrect." Deduct points according to their frequency and severity. Also evaluate grammar errors that significantly hinder comprehension more strictly.

Accuracy

Here you compare the two translations with the original German text to assess whether they match in terms of content. To display the source text, click "Show/hide original text in German" at the top of the page. For each segment, compare which translation is better and click on your preferred one. It will then be highlighted in green. If you find both translations equally good or equally poor, do not mark the segment.

If a translation expresses the exact opposite of the original text, assign at least 1 point, but no more than 2. If information is omitted or invented, deduct points accordingly. Here as well, prefer idiomatic expressions over overly literal translations. If the same information is present but in a different order, this does not need to be penalized as long as the meaning is not changed. Inconsistency in the translation of a particular word—when a word is translated in several different ways—does not have to be penalized, but may be penalized with a maximum deduction of 1 point. Untranslated proper names, provided a translation exists, likewise do not have to be penalized, but may be penalized with a maximum deduction of 1 point.