

Expanding the WMT24++ Benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader

Jannis Vamvas¹ Ignacio Pérez Prat² Not Battesta Soliva¹
Sandra Baltermia-Guetg² Andrina Beeli² Simona Beeli² Madlaina Capeder²
Laura Decurtins² Gian Peder Gregori² Flavia Hobi² Gabriela Holderegger²
Arina Lazzarini² Viviana Lazzarini² Walter Rosselli² Bettina Vital²
Anna Rutkiewicz¹ Rico Sennrich¹

¹University of Zurich ²Lia Rumantscha

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

Abstract

The Romansh language, spoken in Switzerland, has limited resources for machine translation evaluation. In this paper, we present a benchmark for six varieties of Romansh: Rumantsch Grischun, a supra-regional variety, and five regional varieties: Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. Our reference translations were created by human translators based on the WMT24++ benchmark, which ensures parallelism with more than 55 other languages. An automatic evaluation of existing MT systems and LLMs shows that translation out of Romansh into German is handled relatively well for all the varieties, but translation into Romansh is still challenging.

1 Introduction

The automatic evaluation of machine translation (MT) has been widened in recent years to cover more languages and language varieties. While massively multilingual benchmarks such as FLORES (Goyal et al., 2022; NLLB Team et al., 2024) or NTREX (Federmann et al., 2022) include reference translations in hundreds of languages, no dedicated reference translations for the Romansh language have been available so far. In this paper, we close this gap by extending the recent WMT24++ benchmark (Kocmi et al., 2024; Deutsch et al., 2025) with reference translations for six varieties of Romansh, using German as the source language.

There are several reasons why Romansh, which is a language from the Romance family spoken in Switzerland (ISO 639-1: *rm*; ISO 639-2/3: *roh*), has had limited resources for MT evaluation. First, Romansh is considered a minority language, with 40,000–60,000 speakers (Gross, 2004; Grünert, 2024). Secondly, multiple varieties of Romansh need to be considered for a comprehensive evaluation. *Rumantsch Grischun* is a supra-regional variety of the language, often used in official contexts. However, the five regional varieties of Ro-

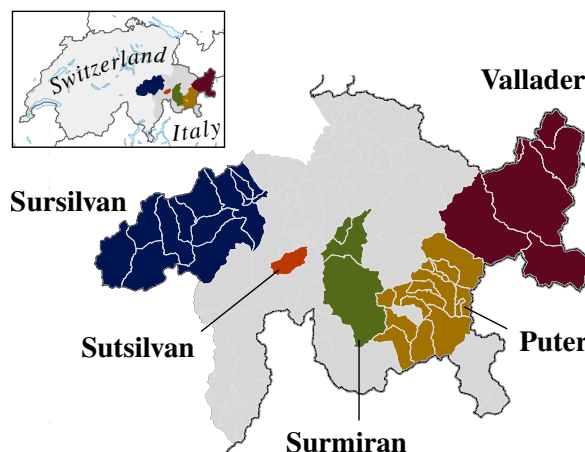


Figure 1: Distribution of Romansh *idioms* (regional varieties) within south-eastern Switzerland. The map shows municipalities where an idiom is officially used in public administration. We extend the WMT24++ benchmark with sets of reference translations for these five idioms, as well as Rumantsch Grischun, a supra-regional variety of Romansh.

mansh (Figure 1), usually referred to as *idioms*, are more widely spoken in everyday life, with limited mutual intelligibility (Gross, 2004).

Prior work on MT for Romansh (Müller et al., 2020; Niklaus et al., 2025) has leveraged multilingual government press releases (Scherrer and Cartoni, 2012), blog posts or federal laws, all of which cover only the Rumantsch Grischun variety. Our benchmark based on WMT24++ enables a more systematic evaluation setup that includes the five idioms and a broader range of domains, such as social media and transcripts of YouTube videos.

We release our benchmark under the Apache 2.0 license.¹ In addition, we use the benchmark to perform a systematic evaluation of MT systems and LLMs on German–Romansh and Romansh–German translation. Results based on automatic evaluation metrics indicate that translation into German achieves reasonable quality for all Romansh

¹<https://hf.co/datasets/ZurichNLP/wmt24pp-rm>

Data Sample	
English (Kocmi et al., 2024)	<i>it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me</i>
German (Deutsch et al., 2025)	<i>Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet</i>
Rumantsch Grischun Code: roh_Latn_ruma1247	<i>Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.</i>
Sursilvan Code: roh_Latn_surs1244	<i>Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.</i>
Surmiran Code: roh_Latn_surm1243	<i>Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir</i>
Sutsilvan Code: roh_Latn_suts1235	<i>Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me</i>
Puter Code: roh_Latn_uppe1396	<i>Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot</i>
Vallader Code: roh_Latn_lowe1386	<i>Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm voul propcha ir per testa</i>

Table 1: Samples of the Romansh varieties that we contribute to the benchmark, plus the English and German segments from prior work. The samples are from the *Social* domain. The language code assigned by the Open Language Data Initiative (OLDI) has three components: the ISO 639-3 language code (roh), the ISO 15924 script code (Latn), and the Glottocode assigned to the variety by Glottolog (Hammarström et al., 2025).

varieties, while translation into Romansh remains challenging, particularly for the less-resourced idioms. Code for reproducing our experiments is available.²

2 Language Overview

2.1 Romansh

Romansh is part of the Romance branch of the Indo-European language family. It is a minority language in the Swiss canton of Graubünden and is treated as one of the country’s four national languages (Grünert, 2018). Its status is considered endangered (Moseley and Nicolas, 2010). Romansh covers an extremely diverse dialect continuum spanning the canton, where roughly 15% of inhabitants speak it as their main language (Gross, 2004). Unlike other dialect continuums, Romansh is not “roofed” by a single standard language (Goebl, 2003). Instead, there are five different written traditions dividing the Romansh-speaking area into regions with their own written standards—known as *idioms*—that differ heavily from each other in all areas of language structure (Liver, 2010; Haiman and Benincà, 1992; Schmid, 1976). In the 1980s,

linguist Heinrich Schmid developed Rumantsch Grischun, a supra-regional, constructed standard, as a *Dachsprache* for Romansh (Muljačić, 2012).

2.2 Sursilvan

Sursilvan is used in the west of Graubünden, in an area mainly covered by the Surselva valley. In many municipalities towards the east, it is still the predominant first language, while German is increasingly dominant moving west (Gross, 2004). Sursilvan is the idiom with the largest population. The written form mainly represents the dialects spoken between Disentis and Ilanz, though the entire area it covers is a continuum exhibiting mutual intelligibility. There were at least 18,000 Sursilvan speakers in the year 2000 (Gross, 2004).

2.3 Sutsilvan

Sutsilvan is spoken in the valley of the Hinterrhein river, though its territory is no longer contiguous. It is the variety with the lowest number of speakers and the highest level of endangerment (Liver, 2014). Large parts of Sutsilvan’s traditional speaker territory became German speaking several centuries ago, and until recently, there was no established written form for Sutsilvan. A concentrated

²https://github.com/ZurichNLP/romansh_mt_eval

effort to change this was initiated by Giuseppe Gan-gale in the 1940s, establishing modern Sutsilvan orthography. His approach, however, sparked debate, and Sutsilvan remains an idiom with hardly any majority Romansh territory (Coray, 2008). There were at least 1,000 Sutsilvan speakers in the year 2000 (Gross, 2004).

2.4 Surmiran

Surmiran is spoken in central Graubünden, namely in the regions of Alvra/Sotses and Surses, the latter being an area where Romansh is still largely present in everyday life (Liver, 2014). Together with Sutsilvan, Surmiran constitutes a bridge between the starkly different dialects of the Surselva region in western Graubünden and the Engadine valley in the east. For this reason, it has previously been suggested as a lingua franca for supra-regional communication (Coray, 2008). Surmiran itself exhibits some peculiarities, however, shared by neither of the two other major Romansh-speaking areas. There were at least 3,000 Surmiran speakers in the year 2000 (Gross, 2004).

2.5 Puter

Puter and Vallader are used as written standards in the Engadine valley, with Puter being used south of Zuoz. The Engadine valley can itself be seen as a continuum of varieties more diverse than the Surselva (Schmid, 1976). Puter reflects characteristics of the dialects in the upper Engadine valley, with more Italian influence than Vallader. Written Puter dates back to 1552 (Obrist, 2022), and is thus the variety with the longest-standing written tradition. Puter is under substantial pressure from German due to growing tourism since the last century (Liver, 2014). Municipalities with a Romansh majority have become scarce. There were at least 5,500 Puter speakers in the year 2000 (Gross, 2004).

2.6 Vallader

Vallader is used in the Lower Engadine valley, north of Zernez, as well as in the Val Müstair. Vallader, unlike Puter, remains a majority language in most of its territory (Liver, 2014). Together with Sursilvan, the Vallader territory represents a stronghold of Romansh. There were at least 6,500 Vallader speakers in the year 2000 (Gross, 2004).

2.7 Rumantsch Grischun

The Rumantsch Grischun variety has a special role in that it is not an idiom, but a written standard devised as a constructed language. It does not reflect any Romansh speaker’s natural speech, but was constructed to be a globally intelligible and neutral written form that could be used to represent Romansh as a language. It was developed by comparing structural and lexical characteristics of the different idioms and determining the most mutually intelligible forms (Schmid, 1982).

Rumantsch Grischun is used for official publications from the canton or the federal government, as well as other institutions addressing the entire Romansh population. More extended promotion of Rumantsch Grischun (including replacing the idioms as the language of literacy at schools) met heavy resistance and caused long-lasting debate (Coray, 2008). Most speakers of Romansh only actively learn their own idiom. Though they may occasionally come into contact with Rumantsch Grischun texts, their knowledge of it is only passive at most.

3 Data Collection

3.1 Choice of Benchmark

We chose to extend the WMT24++ benchmark (Kocmi et al., 2024; Deutsch et al., 2025) based on the following considerations:

- WMT24++ currently covers 55 languages, including the other Swiss national languages (German, French and Italian).
- It is a recent benchmark that is unlikely to suffer from data contamination in LLMs.
- Segments are provided in context, allowing for document-level evaluation.

3.2 Creation of Reference Translations

The data acquisition process was structured into three steps to ensure high quality, consistency, and adherence to idiom-specific conventions.

1. **Translation:** We hired language professionals who are native speakers of both German and the respective Romansh idiom.
2. **Review:** Two expert linguists of Lia Rumantscha reviewed a sample of translations for a representative selection of varieties, and formulated feedback that was communicated to all translators.

↓ pred	gold →					
	RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	764	83	23	43	29	35
Surs.	98	810	67	46	55	54
Suts.	8	12	809	20	6	10
Surm.	19	9	32	811	8	6
Puter	10	12	12	17	648	30
Vall.	61	34	17	23	214	825

Table 2: Confusion matrix of a Romansh language variety classifier when applied to the reference translations.

3. **Revision:** The translators incorporated the feedback into the reference translations.

We provided the translators with a guidelines document, inspired by the WMT24 translator brief (Kocmi et al., 2024). The key points of the guidelines, which we provide in Appendix D, are:

- The German text is the main source for the translation into Romansh, while the English text can be used as an additional reference in case of ambiguity.
- No AI tools should be used for the translation.

The translators and reviewers had access to the complete context of each segment, including a link to the original website from which the segment was extracted (e.g., for segments from the Speech domain, the original YouTube video).

3.3 Challenges in the Data Acquisition

A challenge we encountered in the translation process was that the degree of standardization can vary across text domains. The Romansh idioms are well-standardized, which is reflected in the formal domains *News* and (partially) *Literary*. However, in the *Social* and *Speech* domains, there is more room for individual variation based on the translator’s style or regional background. Therefore, while we consider the reference translations suitable for their intended use of evaluating idiom-aware MT, the dataset does not aim to represent the full spectrum of variation present in the Romansh idioms.

4 Validation Experiments

We perform two automatic validation experiments to confirm that the reference translations are suitable for variety-specific evaluation:

Language Classification We use a fastText classifier (Joulin et al., 2017) trained on a corpus of

↓ sys	ref →					
	RG	Surs.	Suts.	Surm.	Puter	Vall.
RG		60.0	47.5	53.8	47.2	49.4
Surs.	60.7		54.5	49.5	43.1	43.0
Suts.	48.3	54.7		50.9	39.5	39.1
Surm.	54.8	49.8	51.0		43.1	43.3
Puter	47.4	42.7	39.0	42.4		58.7
Vall.	49.3	42.4	38.4	42.4	58.4	

Table 3: Pairwise ChrF scores between the reference translations for the different varieties.

Romansh newspaper articles that were manually labeled with their variety. Table 2 shows that when applied to our reference translations, the classifier predicts the expected variety for the majority of segments, indicating that the reference translations exhibit variety-specific features.

Cross-Variety Scores We calculate pairwise ChrF scores (Popović, 2015) between the sets of references, which are reported in Table 3. The maximum ChrF score across varieties is 60.7 (for Sursilvan–Rumantsch Grischun), which confirms that the sets of reference translations are distinct from each other even for related varieties, allowing for variety-specific evaluation. At the same time, the cross-variety scores are high enough to rule out serious data quality issues, such as a systematic misalignment of segments.

5 Evaluation of MT Systems and LLMs on Translation from and into Romansh

We use our benchmark to evaluate the following machine translation systems and LLMs:

- **MADLAD-400** (Kudugunta et al., 2023), a family of open-source MT models trained on parallel data in more than 450 languages, including Romansh. We report results for the largest, 10.7B-parameter model, using sentence segmentation with SpaCy to translate sentences individually, with a beam size of 5.
- **Supertext**, a commercial MT system that supports German and Romansh, among other languages.³ We use the website of Supertext to translate the segments in an Excel file.
- **Translatur-ia**, a closed, early prototype of an MT system that translates from German into

³<https://supertext.com/>

System	Rumantsch Grischun	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader
MADLAD-400 (10.7B)						
– direct	58.3 / 63.0	52.9 / 54.7	40.6 / 38.1	45.2 / 40.4	49.7 / 49.8	52.8 / 52.9
– pivoting via English	56.1 / 64.9	50.3 / 52.8	39.4 / 37.4	42.4 / 40.1	47.2 / 49.8	49.3 / 51.9
Supertext	72.3 / 92.6	66.9 / 90.7	58.7 / 76.6	62.9 / 81.5	67.0 / 85.2	69.1 / 86.6
Llama 3.3 (70B)	63.1 / 82.8	57.0 / 75.5	48.7 / 59.2	52.1 / 64.3	57.1 / 73.2	60.0 / 75.4
GPT-4o	74.3 / 92.9	70.9 / 92.2	64.2 / 85.2	67.7 / 87.3	71.7 / 90.6	75.1 / 91.1
Gemini 2.5 Flash	75.4 / 93.1	72.1 / 92.9	68.5 / 89.4	71.7 / 90.6	73.5 / 91.7	77.7 / 92.3

Table 4: **Romansh as source language:** ChrF / xCOMET scores of MT systems and LLMs for translation into German from six varieties of Romansh.

System	Rumantsch Grischun	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader
MADLAD-400 (10.7B)						
– direct	48.0	40.7	34.6	37.1	37.0	38.5
– pivoting via English	50.7	43.0	36.1	38.7	38.6	40.1
Translatur-ia	19.7	18.1	16.7	17.4	17.3	17.6
Supertext	68.9	53.2	43.5	47.8	46.7	49.0
Llama 3.3 (70B)	52.1	43.9	36.6	39.3	40.3	42.6
GPT-4o	64.8	60.1	41.4	46.4	52.3	55.9
Gemini 2.5 Flash	66.0	58.7	43.7	50.1	53.8	57.2

Table 5: **Romansh as target language:** ChrF scores of MT systems and LLMs for translation from German into Romansh. Results in gray are based on translations into Rumantsch Grischun, which is the only target variety officially supported by these systems.

Rumantsch Grischun.⁴

- **Llama 3.3** (Grattafiori et al., 2024), an open-source LLM released in November 2024. We use the 70B-parameter version.
- **GPT-4o** (OpenAI et al., 2024), a commercial LLM that was released in May 2024 and was billed at \$2.50 per million input tokens and \$10 per million output tokens.
- **Gemini 2.5 Flash** (Comanici et al., 2025), a commercial LLM that was released in June 2025 and was billed at \$0.30 per million input tokens and \$2.50 per million output tokens. We turn off the ‘thinking’ mode to enable a direct comparison with the other systems.

LLM Prompting When using LLMs for translation, we use the same prompting setup as the WMT24 General Machine Translation Shared Task (Kocmi et al., 2024).⁵ Specifically, we use

⁴<https://translaturia.fhgr.ch/>

⁵<https://github.com/wmt-conference/wmt-collect-translations>

3-shot prompting with temperature set to zero. The prompt template is listed in Appendix B.⁶ As few-shot examples, we use typical example sentences from the fable *The Fox and the Crow* (Gross, 2004), which we list in Appendix C.

Quality Metrics For evaluating translation quality, we use ChrF (Popović, 2015), a metric based on character n-grams that does not require word segmentation, via SacreBLEU (Post, 2018).⁷

For evaluating translations from Romansh into German, we additionally use xCOMET (Guerreiro et al., 2024), a neural metric that was ranked highly in the WMT24 Metrics Shared Task (Freitag et al., 2024). We use model version XCOMET-XL⁸ in the *reference-only* mode, i.e., we do not provide the Romansh source sequence to the metric, a language it

⁶A limitation of this prompt template is that it does not provide the LLM with context beyond the segment that is currently being translated. We opt to keep the setup similar to WMT24 and leave document-level evaluation to future work.

⁷Signature:

#:1|c:mixed|e:yes|nc:6|nw:0|s:no|v:2.5.1

⁸<https://hf.co/Unbabel/XCOMET-XL>

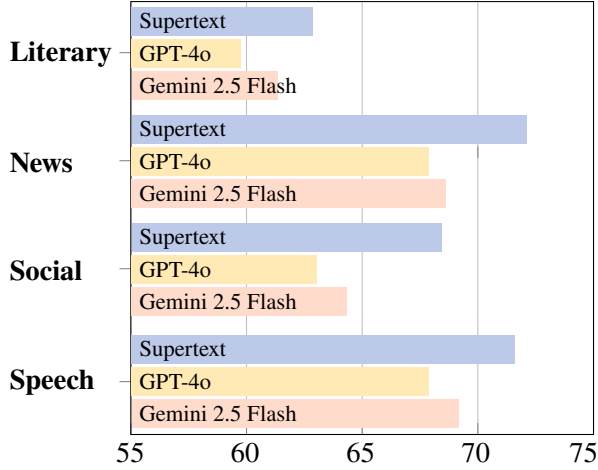


Figure 2: Domain-specific ChrF scores of systems translating from German into Rumantsch Grischun.

has not seen during training. While xCOMET is designed to support such monolingual, *reference-only* evaluation, this mode has not been as extensively validated as other modes. Thus, xCOMET complements ChrF but should be interpreted with some caution. Following Kocmi et al. (2024), we report xCOMET as the macro-average over domains to control for different segment granularities.

6 Results

6.1 Comparison of Translation Directions

Table 4 shows the results for translation from Romansh into German, while Table 5 shows the results for translation from German into Romansh. The former performs consistently better than the latter; this is observed both for the supervised MT systems (MADLAD, Supertext) and for the LLMs.

Comparing performance across the six varieties of Romansh, we find that translation out of Romansh into German is relatively robust to linguistic variation: For Gemini, the gap in terms of ChrF between the minimum and maximum is $77.7 - 68.5$. In contrast, for translation into Romansh, the gap is $66.0 - 43.7$. Future work could exploit this asymmetry by using back-translation (Sennrich et al., 2016) for augmenting monolingual Romansh text with synthetic German translations.

6.2 Ranking of Models

For translation from Romansh into German, we report both ChrF and xCOMET scores in Table 4. We find that the system rankings are largely consistent between the two metrics, on average over the four domains, with Gemini 2.5 Flash achiev-

↓ tgt	ref →					
	RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	66.0	51.6	42.6	46.7	46.2	48.5
Surs.	57.0	58.7	42.6	44.8	43.7	45.1
Suts.	56.7	50.3	43.7	46.6	44.1	45.8
Surm.	52.3	46.5	42.5	50.1	43.8	44.5
Puter	47.5	42.1	38.0	41.2	53.8	54.6
Vall.	49.9	43.2	38.6	41.9	52.7	57.2

Figure 3: “Confusion matrix” of Gemini 2.5 Flash when translating into specific Romansh varieties. To visualize the degree to which the LLM output matches the requested variety (**tgt**), we evaluate the outputs with each set of reference translations (**ref**). A system that adheres to the requested target variety will achieve higher ChrF scores in the diagonal cells than in the off-diagonal cells.

ing the highest scores according to both metrics. MADLAD-400 underperforms the other systems, likely due to the limited Romansh training data and the massively multilingual nature of the model.

In the German–Romansh direction, where ChrF is the only available metric (Table 5), we find that Gemini 2.5 Flash again achieves the highest scores for four out of six varieties. Supertext is the highest-ranked system for translation into Rumantsch Grischun, which is the officially supported target variety of this product.

6.3 Domain Difficulty

Figure 2 compares ChrF scores for German–Rumantsch Grischun translation for the four domains covered by the WMT24++ benchmark. The figure indicates that the *News* domain is the least challenging for all systems, which is consistent with findings of the WMT24 task for other languages (Kocmi et al., 2024). Surprisingly, lowest ChrF scores are achieved in the *Literary* domain, while the human evaluation of the WMT24 task did not find a systematic difference between the *Literary* and *News* domains in terms of difficulty. The *Speech* domain yields similar scores to *News*, and *Social* is slightly more challenging. Detailed results for each domain and variety (Appendices E and F) indicate that this pattern is consistent across varieties.

6.4 Target Variety Adherence of LLMs

While MADLAD and Supertext are limited to Rumantsch Grischun as the target variety, the LLMs can be prompted to produce translations in any of the six varieties. This raises the question of

whether the LLMs actually adhere to the requested target variety. Figure 3 shows a “confusion matrix” for Gemini 2.5 Flash, where we evaluate the system output not only with the reference translations for the requested target variety, but also with contrastive reference translations for the other varieties. The results suggest that state-of-the-art LLMs already have some degree of idiom awareness, but gravitate towards the higher-resource varieties (Rumantsch Grischun, Sursilvan, and Vallader).

7 Conclusion

This work fills a long-standing gap in the evaluation of machine translation for the Romansh language: the creation of a benchmark for the six main varieties of Romansh, and the provision of baseline results for existing MT systems and LLMs that cover Romansh.

Acknowledgements

We thank RTR and Fundaziun Patrimoni Cultural RTR for their support. We are grateful to Kirill Semenov for assistance with the poster presentation, Zachary Hopton for assistance with editing the manuscript, Noëmi Aepli and Martin Volk for helpful advice, and Tom Kocmi for help with the WMT24 benchmark.

Figure 1 uses maps from the [Canton of Grisons](#) and from [Wikimedia Commons](#) (User:Tomchen1989, User:NordNordWest, User:TUBS, CC BY-SA 3.0).

Author Contributions

JV: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft.

IPP: Funding acquisition, Project administration, Supervision, Writing – review & editing.

NBS: Writing – original draft.

SBG, AB, SB, MC, GPG, FH, GH, AL, VL, WR: Translation and/or translation quality assurance.

LD, BV: Linguistic supervision of translation workflow.

AR: Software.

RS: Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

References

- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, 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 3284 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.
- 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.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chikui Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.
- Hans Goebel. 2003. Externe Sprachgeschichte der romanischen Sprachen im Zentral-und Ostalpenraum. *Romanische Sprachgeschichte: Ein internationales Handbuch zur Geschichte der romanischen Sprachen*, 1:747–773.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

- Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Manfred Gross. 2004. [Romansh: Facts & Figures](#), [2nd rev. and updated ed.] edition. Lia rumantscha, Chur.
- Matthias Grünert. 2018. [Multilingualism in Switzerland](#). In *Manual of Romance Sociolinguistics*, pages 526–548. De Gruyter. Section: Manual of Romance Sociolinguistics.
- 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.
- John Haiman and Paola Benincà. 1992. [The rhaeto-romance languages](#). Routledge, London.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. [glottolog/glottolog: Glottolog database 5.2.1](#).
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. 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.
- 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.
- Ricarda Liver. 2010. *Rätoromanisch: eine Einführung in das Bündnerromanische*, 2., überarbeitete und erweiterte Auflage edition. Narr Studienbücher. Narr Verlag, Tübingen.
- Ricarda Liver. 2014. [Le romanche des Grisons](#). In *Manuel des langues romanes*, pages 413–446. De Gruyter. Section: Manuel des langues romanes.
- Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the world’s languages in danger*, 3rd ed., entirely revised, enlarged and updated edition. UNESCO, Paris. Series: Memory of peoples series Book Title: Atlas of the world’s languages in danger.
- Žarko Muljačić. 2012. [Über den Begriff Dachsprache](#). In Ulrich Ammon, editor, *Status and Function of Languages and Language Varieties*, pages 256–277. De Gruyter.
- 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 Gösen, 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.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, and 20 others. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Philipp Ombert. 2022. [Discourse traditions in the history of Romansh](#). In *Manual of Discourse Traditions in Romance*, pages 615–632. De Gruyter. Section: Manual of Discourse Traditions in Romance.
- OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- 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.

Yves Scherrer and Bruno Cartoni. 2012. [The trilingual ALLEGRA corpus: Presentation and possible use for lexicon induction](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2890–2896, Istanbul, Turkey. European Language Resources Association (ELRA).

Heinrich Schmid. 1976. [Zur Gliederung des Bündnerromanischen](#). *Annalas da la Societad Retorumantscha*, 89:7–62. Publisher: Stampa Romontscha.

Heinrich Schmid. 1982. *Richtlinien für die Gestaltung einer gesamtbündnerromanischen Schriftsprache: Rumantsch grischun*, [2. aufl.] edition. Lia Rumantscha, Cuira.

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.

A Dataset Statistics

Variety	Segments					Tokens				
	Lit.	News	Soc.	Speech	Total	Lit.	News	Soc.	Speech	Total
German (Deutsch et al., 2025)	206	149	531	111	998	9 814	9 416	10 556	10 157	39 954
RG	206	149	531	111	998	10 893	11 626	11 949	10 174	44 653
Sursilvan	206	149	531	111	998	10 609	11 287	11 913	10 144	43 964
Sutsilvan	206	149	531	111	998	11 147	11 667	12 241	10 378	45 444
Surmiran	206	149	531	111	998	10 718	11 424	11 907	10 167	44 227
Puter	206	149	531	111	998	11 325	11 595	12 267	10 490	45 688
Vallader	206	149	531	111	998	11 412	11 700	12 330	10 513	45 966

Table 6: Dataset statistics for each language variety. Token counts are based on whitespace tokenization.

B Prompt Template for LLMs

The below example is parameterized as follows:

- Source language: German
- Target language: Romansh
- Target variety: Rumantsch Grischun
- Source sentence: “*Sisos Darstellungen von Land und Wasser in neuer Ausstellung*”

Translate the following segment surrounded in triple backticks into Romansh (Rumantsch Grischun variety). The German segment:

```
```Sisos Darstellungen von Land und Wasser in neuer Ausstellung```
```

## C Few-shot Examples for LLM Prompting

### German:

Der Fuchs war wieder einmal hungrig.

Da sah er auf einer Tanne einen Raben, der ein Stück Käse in seinem Schnabel hielt.

Das würde mir schmecken, dachte er, und rief dem Raben zu: «Wie schön du bist! Wenn dein Gesang ebenso schön ist wie dein Aussehen, dann bist du der Schönste von allen Vögeln».

### English:

The fox was hungry yet again.

There, he saw a raven upon a fir holding a piece of cheese in its beak.

This I would like, he thought, and shouted at the raven: "You are so beautiful! If your singing is as beautiful as your looks, then you are the most beautiful of all birds."

### Rumantsch Grischun:

La vulp era puspè ina giada fomentada.

Qua ha ella vis sin in pign in corv che tegneva in toc chaschiel en ses pichel.

Quai ma gustass, ha ella pensà, ed ha clamà al corv: «Tge bel che ti es! Sche tes chant è uschè bel sco tia parita, lur es ti il pli bel utschè da tuts».

### Sursilvan:

L'uolp era puspei inagada fomentada.

Cheu ha ella viu sin in pegn in tgaper che teneva in toc caschiel en siu bec.

Quei gustass a mi, ha ella tertgau, ed ha clamau al tgaper: «Tgei bi che ti eis! Sche tiu cant ei aschi bials sco tia cumparsa, lu eis ti il pli bi utschi da tuts».

### Sutsilvan:

La gualp eara puspe egn'eada fumantada.

Qua â ella vieu sen egn pegn egn corv ca taneva egn toc caschiel ainten sieus pecel.

Quegl gustass a mei, â ella tartgieu, ed ha clamo agli corv: «Tge beal ca tei es! Scha tieus tgànt e aschi beal sco tia pareta, alura es tei igl ple beal utschi da tuts».

### Surmiran:

La golp era puspe eneda famantada.

Co ò ella via sen en pegn en corv tgi tigniva in toc caschiel an sies pecal.

Chegl am gustess, ò ella panso, ed ò clamo agl corv: «Tge bel tgi te ist! Schi ties cant è schi bel scu tia parentscha, alloura ist te igl pi bel utschel da tots».

### Puter:

La vuolp d'eira darcho üna vouta famantada.

Co ho'la vis sün ün pin ün corv chi tгнаiva ün töch chaschöl in sieu pical.

Que am gustess, ho'la penso, ed ho clamo al corv: «Che bel cha tü est! Scha tieu chaunt es uschè bel scu tia apparentscha, alura est tü il pü bel utschè da tuots».

### Vallader:

La vuolp d'eira darcheu üna jada fomantada.

Qua ha'la vis sün ün pin ün corv chi tгнаiva ün toc chaschöl in seis pical.

Quai am gustess, ha'la pensà, ed ha clomà al corv: «Che bel cha tü est! Scha teis chant es uschè bel sco tia apparentscha, lura est tü il plü bel utschè da tuots».

## D Translation Guidelines (in German)

### 1. Wofür werden die Übersetzungen benötigt?

Wir verwenden die von Ihnen erstellten Übersetzungen für die Evaluierung von maschinellen Übersetzungssystemen. Der Output der Übersetzungssysteme wird mit Ihrer Übersetzung verglichen – je ähnlicher der Output, desto besser das Übersetzungssystem.

Die Referenzübersetzungen werden nicht für das Training der Übersetzungssysteme verwendet.

### 2. Wie wurden die Texte ausgewählt?

Die Texte stammen aus einem bestehenden Datensatz («WMT24»), der zuvor bereits aus dem Englischen in 55 verschiedene Sprachen übersetzt worden ist, darunter Deutsch.

Der Datensatz setzt sich aus vier Textsorten zusammen:

- **literary:** Fan-Fiction, welche auf der Website «Archive of Our Own» veröffentlicht wurde.
- **news:** Zufällig ausgewählte Online-News vom Januar 2024.
- **social:** Zufällig ausgewählte Threads aus dem Sozialen Netzwerk «Mastodon».
- **speech:** Transkripte zufällig ausgewählter YouTube-Videos.

### 3. Wie ist die Excel-Datei aufgebaut?

Wir erstellen für jedes Idiom eine eigene Excel-Datei. Die Datei enthält vier Tabellen für die vier Textsorten, und jede Zeile in der Tabelle entspricht einem Textsegment. Mehrere Textsegmente setzen sich zu zusammenhängenden Dokumenten zusammen, welche durch graue Leerzeilen voneinander abgetrennt sind.

Die Tabellen sind wie folgt aufgebaut:

- **English:** Originaler Text auf Englisch.
- **document\_id:** ID des Dokuments.
- **segment\_id:** ID des Segments.
- **url:** Webseite, von welcher das Dokument ursprünglich bezogen wurde. Kann optional aufgerufen werden, um den Kontext des Dokuments nachzuvollziehen.
- **German:** Referenzübersetzung auf Deutsch.
- **translation:** Hier soll die Übersetzung auf Rätoromanisch eingetragen werden.
- **comment:** Kann von Ihnen optional benutzt werden, um uns einen wichtigen Kommentar zu hinterlassen.

### 4. Welche Anforderungen gelten an die Übersetzungen?

- Die Referenzübersetzungen repräsentieren Ihre Erwartungen an den Output eines guten maschinellen Übersetzungssystems. Überlegen Sie sich: Was für eine Übersetzung würde ein gutes Übersetzungssystem (im Stil von DeepL oder Google Translate) für das entsprechende romanische Idiom erzeugen?
- Massgeblich für Ihre Übersetzung sollte der Text auf Deutsch sein. Der englische Text kann optional zu Rate gezogen werden, falls der deutsche Text mehrdeutig ist.
- Weil wir Ihre Referenzübersetzungen für die Evaluierung verwenden möchten, bitten wir Sie, keine AI-Tools zu verwenden. Bitte erstellen Sie die Übersetzungen von Grund auf in Ihren eigenen Worten.



- Erlaubte Hilfsmittel: Wörterbücher, Translation Memories, ...
- Nicht erlaubte Hilfsmittel: ChatGPT, Copilot, Supertext/Textshuttle, DeepL, ...

Wir können den Übersetzungen ansehen, wenn sie mit AI-Unterstützung erstellt wurden, und wir müssten im schlimmsten Fall die Übersetzungen noch einmal neu erstellen lassen (dies ist leider in der Vergangenheit schon vorgekommen).

- Bitte fügen Sie keine Übersetzungsalternativen oder Erklärungen in Klammern in die Übersetzungen ein.
- Bitte übersetzen Sie die Information in den Texten vollständig.
- Lehnwörter, Eigennamen etc. dürfen gerne auf Englisch oder Deutsch belassen werden, falls ein gutes rätoromanisches Übersetzungssystem das gleiche machen sollte.
- Betreffend die Textsorte «social»:
  - Viele Posts enthalten Fachbegriffe, Slang, Abkürzungen. Durch die Übersetzung vom Englischen ins Deutsche dürfte vieles schon weniger kryptisch geworden sein. Falls dennoch bei einem Post Unsicherheit besteht, können Ihnen diese Notizen weiterhelfen: <https://github.com/wmt-conference/wmt24-news-systems/blob/main/README-social-domain-translation-notes.pdf>
  - Nutzernamen wurden anonymisiert (z.B. @user1, @user2). Bitte übernehmen Sie die Nutzernamen eins zu eins, d.h. Sie müssen diese nicht übersetzen.
  - Hingegen dürfen #Hashtags gerne übersetzt werden, falls dies im Kontext Sinn macht.

## E Detailed Results for RM-DE

### E.1 Rumantsch Grischun to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	50.6 / 52.8	67.9 / 81.3	53.0 / 71.7	59.5 / 46.4	57.8 / 63.0
– pivoting via English	49.8 / 57.1	63.2 / 79.1	51.9 / 73.1	58.0 / 50.2	55.7 / 64.9
Supertext	68.3 / 89.9	71.1 / 95.1	73.9 / <b>96.2</b>	75.8 / 89.1	72.3 / 92.6
Llama 3.3 (70B)	56.8 / 75.3	68.0 / 91.3	58.2 / 84.3	68.6 / 80.4	62.9 / 82.8
GPT-4o	<b>71.5 / 91.8</b>	74.1 / <b>95.9</b>	73.6 / 94.3	78.3 / 89.6	74.4 / 92.9
Gemini 2.5 Flash	<b>71.5 / 90.6</b>	<b>74.4 / 95.7</b>	<b>76.5 / 95.9</b>	<b>79.5 / 90.2</b>	<b>75.5 / 93.1</b>

Table 7: ChrF / xCOMET scores for translation from Rumantsch Grischun into German.

### E.2 Sursilvan to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	45.7 / 44.0	64.4 / 76.6	46.9 / 65.8	52.2 / 32.5	52.3 / 54.7
– pivoting via English	44.0 / 45.0	59.1 / 70.2	45.2 / 65.8	51.1 / 30.4	49.9 / 52.8
Supertext	63.6 / 87.1	68.7 / 94.8	64.5 / 94.4	70.6 / 86.4	66.9 / 90.7
Llama 3.3 (70B)	50.5 / 63.7	64.3 / 88.4	50.4 / 77.0	61.7 / 72.9	56.7 / 75.5
GPT-4o	68.6 / <b>90.7</b>	72.2 / 95.5	67.8 / 93.0	<b>74.9 / 89.5</b>	70.9 / 92.2
Gemini 2.5 Flash	<b>69.7 / 90.5</b>	<b>72.9 / 95.7</b>	<b>70.8 / 95.7</b>	<b>74.9 / 89.5</b>	<b>72.1 / 92.9</b>

Table 8: ChrF / xCOMET scores for translation from Sursilvan into German.

### E.3 Sutsilvan to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	33.1 / 31.4	54.1 / 50.3	32.8 / 48.4	40.0 / 22.3	40.0 / 38.1
– pivoting via English	32.9 / 31.5	50.3 / 47.2	33.1 / 49.8	39.4 / 21.2	38.9 / 37.4
Supertext	55.1 / 67.9	62.9 / 87.8	54.7 / 86.1	61.1 / 64.8	58.5 / 76.6
Llama 3.3 (70B)	43.0 / 44.3	57.6 / 74.8	40.3 / 61.8	52.9 / 55.7	48.4 / 59.2
GPT-4o	61.5 / 80.5	70.0 / 93.3	54.4 / 81.6	70.3 / 85.2	64.1 / 85.2
Gemini 2.5 Flash	<b>65.7 / 85.0</b>	<b>71.2 / 94.2</b>	<b>64.6 / 92.4</b>	<b>72.3 / 86.1</b>	<b>68.5 / 89.4</b>

Table 9: ChrF / xCOMET scores for translation from Sutsilvan into German.

#### E.4 Surmiran to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	37.8 / 34.1	58.6 / 53.0	37.9 / 52.7	43.9 / 21.7	44.5 / 40.4
– pivoting via English	36.9 / 36.1	53.0 / 50.4	36.0 / 50.6	42.0 / 23.1	41.9 / 40.1
Supertext	59.6 / 76.9	67.8 / 91.6	58.9 / 88.6	64.4 / 69.0	62.7 / 81.5
Llama 3.3 (70B)	46.7 / 52.9	60.9 / 78.7	44.2 / 66.0	55.2 / 59.7	51.7 / 64.3
GPT-4o	66.4 / 86.2	71.6 / 92.3	61.3 / 85.6	71.5 / 85.1	67.7 / 87.3
Gemini 2.5 Flash	<b>69.3 / 88.5</b>	<b>72.8 / 94.3</b>	<b>69.6 / 93.0</b>	<b>75.2 / 86.7</b>	<b>71.7 / 90.6</b>

Table 10: ChrF / xCOMET scores for translation from Surmiran into German.

#### E.5 Puter to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	42.6 / 41.0	63.3 / 68.7	41.8 / 58.6	48.2 / 30.9	49.0 / 49.8
– pivoting via English	41.9 / 42.9	57.1 / 63.2	40.4 / 60.2	47.4 / 32.9	46.7 / 49.8
Supertext	63.9 / 81.6	71.3 / 93.4	62.6 / 89.4	69.5 / 76.4	66.8 / 85.2
Llama 3.3 (70B)	52.5 / 63.6	64.6 / 84.8	49.0 / 73.8	61.0 / 70.4	56.8 / 73.2
GPT-4o	69.7 / 88.7	73.6 / 94.0	67.0 / 91.7	76.6 / <b>88.0</b>	71.7 / 90.6
Gemini 2.5 Flash	<b>71.5 / 89.7</b>	<b>74.6 / 94.6</b>	<b>70.7 / 94.5</b>	<b>77.1 / 88.0</b>	<b>73.5 / 91.7</b>

Table 11: ChrF / xCOMET scores for translation from Puter into German.

#### E.6 Vallader to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	45.9 / 46.9	65.0 / 71.5	46.1 / 63.8	51.7 / 29.3	52.2 / 52.9
– pivoting via English	43.1 / 47.8	59.6 / 65.7	43.7 / 64.3	48.7 / 29.7	48.8 / 51.9
Supertext	65.6 / 83.3	71.6 / 92.5	66.7 / 91.7	72.1 / 78.9	69.0 / 86.6
Llama 3.3 (70B)	54.7 / 67.2	67.3 / 86.2	52.8 / 78.6	64.1 / 69.5	59.8 / 75.4
GPT-4o	72.4 / 90.1	77.3 / 94.5	71.2 / 91.2	79.4 / <b>88.4</b>	75.1 / 91.1
Gemini 2.5 Flash	<b>74.7 / 91.1</b>	<b>78.9 / 94.7</b>	<b>76.1 / 95.3</b>	<b>80.9 / 88.1</b>	<b>77.6 / 92.3</b>

Table 12: ChrF / xCOMET scores for translation from Vallader into German.

## F Detailed Results for DE–RM

### F.1 German to Rumantsch Grischun

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	42.9	56.3	42.6	48.6	47.6
– pivoting via English	45.6	56.6	47.6	51.8	50.4
Translatur-ia	17.1	20.1	23.4	17.2	19.5
Supertext	<b>62.8</b>	<b>72.1</b>	<b>68.4</b>	<b>71.6</b>	<b>68.7</b>
Llama 3.3 (70B)	46.8	57.7	48.6	54.0	51.8
GPT-4o	59.9	67.9	63.0	67.9	64.7
Gemini 2.5 Flash	61.3	68.6	64.3	69.2	65.8

Table 13: ChrF scores for translation from German into Rumantsch Grischun.

### F.2 German to Sursilvan

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	39.7	50.7	39.4	44.5	43.6
GPT-4o	<b>56.2</b>	<b>63.7</b>	<b>57.5</b>	<b>62.4</b>	<b>59.9</b>
Gemini 2.5 Flash	55.4	61.8	56.4	60.6	58.5

Table 14: ChrF scores for translation from German into Sursilvan.

### F.3 German to Sutsilvan

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	32.4	43.0	33.0	36.8	36.3
GPT-4o	37.1	47.5	39.6	40.1	41.1
Gemini 2.5 Flash	<b>40.5</b>	48.9	39.8	<b>44.5</b>	<b>43.4</b>

Table 15: ChrF scores for translation from German into Sutsilvan.

### F.4 German to Surmiran

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	34.7	46.5	35.5	38.8	38.9
GPT-4o	42.2	53.5	43.1	44.8	45.9
Gemini 2.5 Flash	<b>46.9</b>	<b>55.7</b>	<b>47.0</b>	<b>49.6</b>	<b>49.8</b>

Table 16: ChrF scores for translation from German into Surmiran.

## F.5 German to Puter

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	35.8	47.0	36.0	41.0	39.9
GPT-4o	48.0	57.5	<b>49.2</b>	53.4	52.0
Gemini 2.5 Flash	<b>51.3</b>	<b>59.9</b>	48.5	<b>54.8</b>	<b>53.6</b>

Table 17: ChrF scores for translation from German into Puter.

## F.6 German to Vallader

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	38.3	49.2	38.6	42.9	42.2
GPT-4o	52.7	60.2	53.2	56.5	55.7
Gemini 2.5 Flash	<b>53.8</b>	<b>61.6</b>	<b>54.2</b>	<b>58.2</b>	<b>57.0</b>

Table 18: ChrF scores for translation from German into Vallader.



## **G Examples for System Outputs**

### **G.1 German to Romansh**

[English:] *it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me*

German source: *Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet*

#### **Rumantsch Grischun**

Reference: *Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.*

MADLAD-400 (10.7B) direct: *Sembla ch'iMessage na funcziunia betg sur WiFi, quai che n'ha nagin sens per mai*

Translatur-ia: *tuttinascha fitg lungascha betg pitschnas la kienì sur la qualafizistgeli, ma prognostitgà*

Supertext: *Para na funcziuna gnanc iMessage sur WiFi, quai che ma resplenda betg dal tut*

Llama 3.3 (70B): *Apparentamain funcziuna schiz ni iMessage sur WiFi, quai ch'ha a mai betg dal tut enclaus.*

GPT-4o: *Apparentamain na funcziuna gnanc iMessage sur WiFi, quai che na ma para betg del tut cler*

Gemini 2.5 Flash: *Apparentamain na funcziunescha gnanc iMessage sur WiFi, quai che na ma para betg dal tutafatg cler.*

#### **Sursilvan**

Reference: *Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.*

Llama 3.3 (70B): *Scheppa che sch'iva nean iMessage sur WiFi, quai che mi n'entuppa betg dal tut*

GPT-4o: *Apparentamein funcziunescha buc ne iMessage sur WiFi, quei cheu ei buc clar a mi*

Gemini 2.5 Flash: *Apparentamein funcziunescha gnanc iMessage sur WiFi, quei che jeu capeschel buca dal tut.*

#### **Sutsilvan**

Reference: *Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me*

Llama 3.3 (70B): *Apparentamain funcziuna schi betg iMessage sur WiFi, sche mi nu cumpreta betg del tut*

GPT-4o: *Apparentamein fonziunescha gnanc iMessage sur WiFi, tgei ca mei betg clarifica cupleinamein.*

Gemini 2.5 Flash: *Apparentamagn na funcziunescha betg gnieu iMessage sur WiFi, tgei ca na clerescha betg dal tut a mei.*

#### **Surmiran**

Reference: *Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir*

Llama 3.3 (70B): *Apparentamain funcziuna betg ni ir iMessage sur WiFi, sche che m'en betg del tut clar*

GPT-4o: *Apparentamaint funcziunescha gnanc iMessage sur WiFi, tge nun è dal tut cler per mai.*

Gemini 2.5 Flash: *Apparentamaintg na funcziunescha betg gniev iMessage sur WiFi, chegl tgi na m'è betg dal tutafatg cler.*

#### **Puter**

Reference: *Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot*

Llama 3.3 (70B): *Apparentamain funcziuna betg era neir iMessage sur WiFi, cha m'ha betg dal tut en persvasiun*

GPT-4o: *A para cha iMessage nu funcziuna nemain sur WiFi, quai nu'm para fich cler*

Gemini 2.5 Flash: *Apparentamaing nu funcziunescha neir betg iMessage sur WiFi, che nu'm clerescha*

*dal tuot.*

### **Vallader**

Reference: *Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm vould propcha ir per testa*

Llama 3.3 (70B): *Apparentamain funcziunia neir il iMessage sur WiFi, quai ch'ha mì nua betg dal tut en la testa*

GPT-4o: *Apparentamaing nu funcziuna neir iMessage sur WiFi, quai nu'm para fich cler.*

Gemini 2.5 Flash: *Apparentamaing nu funcziunescha neir betg iMessage sur WiFi, che nu'm para dal tuot cler.*

### **G.2 Romansh to German**

[English:] *it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me*

German reference: *Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet*

### **Rumantsch Grischun**

Romansh source: *Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.*

MADLAD-400 (10.7B) direct: *Damit iMessage über WLAN funktioniert, Ich weiß nicht, ob ich das jemals wiederholen werde.*

Supertext: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was für mich nicht ganz selbstverständlich ist.*

Llama 3.3 (70B): *Dass nicht einmal iMessage über WiFi funktioniert, ist überhaupt nicht offensichtlich.*

GPT-4o: *Es scheint, dass nicht einmal iMessage über WiFi funktioniert, was für mich nicht ganz nachvollziehbar ist.*

Gemini 2.5 Flash: *Es scheint, dass nicht einmal iMessage über WLAN funktioniert, was für mich nicht ganz offensichtlich ist.*

### **Sursilvan**

Romansh source: *Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.*

MADLAD-400 (10.7B) direct: *Er stellt fest, dass die Wii-Fernbedienung nicht funktioniert, und fragt sich, was er damit anfangen soll.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was ich wirklich nicht verstehen kann.*

Llama 3.3 (70B): *Weil ich als Funktion habe, dass ich iMessage über WiFi nutzen kann, was ich aber nicht richtig verstehe.*

GPT-4o: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was ich wirklich nicht verstehen kann.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage nicht einmal über WLAN, was ich nicht wirklich verstehen kann.*

### **Sutsilvan**

Romansh source: *Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me*

MADLAD-400 (10.7B) direct: *Die erste Singleauskopplung des Debütalbums, I Don't Want to Miss a Thing, erreichte Platz 1 der US-Charts.*

Supertext: *Anscheinend funktioniert nicht einmal iMessage über WiFi, was auf jeden Fall nicht*

*funktioniert.*

Llama 3.3 (70B): *Damit Funktionen wie iMessage über das WiFi funktionieren, was aber nicht gerade mein Ding ist*

GPT-4o: *Sogar iMessage funktioniert nicht über das WLAN, was für mich keinen Sinn ergibt.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage nicht mal über WLAN, was für mich nicht wirklich Sinn macht*

### **Surmiran**

Romansh source: *Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir*

MADLAD-400 (10.7B) direct: 2000: *1. Preis beim Wettbewerb für Jugendtheaterstücke der Sächsischen Staatsoper Unter den Linden* 2001: *1. Preis beim Wettbewerb für Jugendtheaterstücke der Sächsischen Staatsoper.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was man wirklich nicht verstehen kann.*

Llama 3.3 (70B): *Um Funktionen wie iMessage über WiFi zu nutzen, muss man wissen, dass das nicht ganz umsonst ist*

GPT-4o: *Anscheinend funktioniert iMessage nicht über WiFi, das kann ich nicht wirklich verstehen.*

Gemini 2.5 Flash: *Anscheinend funktioniert mein iMessage nicht über WLAN, was ich nicht wirklich verstehen kann.*

### **Puter**

Romansh source: *Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot*

MADLAD-400 (10.7B) direct: *Ich weiß nicht, ob iMessage über WLAN funktioniert, was ich nicht verstehe.*

Supertext: *Auch wenn iMessage über WiFi nicht funktioniert, ist das nicht ganz nachvollziehbar.*

Llama 3.3 (70B): *Denn nun funktioniert auch iMessage nicht mehr über WiFi, weil ich nicht mehr online bin*

GPT-4o: *Es scheint, dass iMessage nicht einmal über WiFi funktioniert, was ich überhaupt nicht verstehe.*

Gemini 2.5 Flash: *Aber dass iMessage auch über WLAN nicht funktioniert, das verstehe ich überhaupt nicht.*

### **Vallader**

Romansh source: *Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm voul propcha ir per testa*

MADLAD-400 (10.7B) direct: *Anscheinend funktioniert iMessage nicht einmal mehr über WLAN, was mich sehr ärgert.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was mir wirklich nicht einleuchtet.*

Llama 3.3 (70B): *Aber offensichtlich funktioniert nicht einmal iMessage über WiFi, was mich nicht weiter wundert*

GPT-4o: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was mir wirklich nicht in den Kopf will.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage über WLAN nicht einmal, was mir wirklich nicht in den Kopf will.*