Bringing Ladin to FLORES+

Samuel Frontull[1], Thomas Ströhle[1], Carlo Zoli[2], Werner Pescosta[3], Ulrike Frenademez[3], Matteo Ruggeri[3], Daria Valentin[3], Karin Comploj[3], Gabriel Perathoner[3], Silvia Liotto[3], Paolo Anvidalfarei[3]

[1]Department of Computer Science, University of Innsbruck, Austria
[2]Faculty of Education, Free University of Bozen-Bolzano, Italy
[3]Ladin Cultural Institute "Micurá de Rü", San Martin de Tor, Italy

Correspondence: samuel.frontull@uibk.ac.at

Abstract

Recent advances in neural machine translation (NMT) have opened new possibilities for developing translation systems also for smaller, so-called low-resource, languages. The rise of large language models (LLMs) has further revolutionized machine translation by enabling more flexible and context-aware generation. However, many challenges remain for low-resource languages, and the availability of high-quality, validated test data is essential to support meaningful development, evaluation, and comparison of translation systems. In this work, we present an extension of the FLORES+ dataset for two Ladin variants, Val Badia and Gherdëina, as a submission to the Open Language Data Initiative Shared Task 2025. To complement existing resources, we additionally release two parallel datasets for Gherdëina–Val Badia and Gherdëina–Italian. We validate these datasets by evaluating state-of-the-art LLMs and NMT systems on this test data, both with and without leveraging the newly released parallel data for fine-tuning and prompting. The results highlight the considerable potential for improving translation quality in Ladin, while also underscoring the need for further research and resource development, for which this contribution provides a basis.

1 Introduction

In recent years, the field of machine translation (MT) and natural language processing has advanced rapidly and the transformer-based models played a key role in this development (Vaswani et al., 2023; Aharoni et al., 2019). This paradigm shift has enabled the development of high-quality MT systems for major languages as well as the adaptation of such systems to low-resource languages by leveraging pre-trained knowledge (Zoph et al., 2016; Kocmi and Bojar, 2018). Earlier rule-based and statistical MT approaches lacked this capability as they depended on large, clean, and domain-specific parallel corpora, thus limiting the availability of these technologies for most low-resource languages. The release of large-scale multilingual systems, such as No Language Left Behind (NLLB) (NLLB Team et al., 2024), has demonstrated that scalable and accurate translation is possible even for smaller languages.

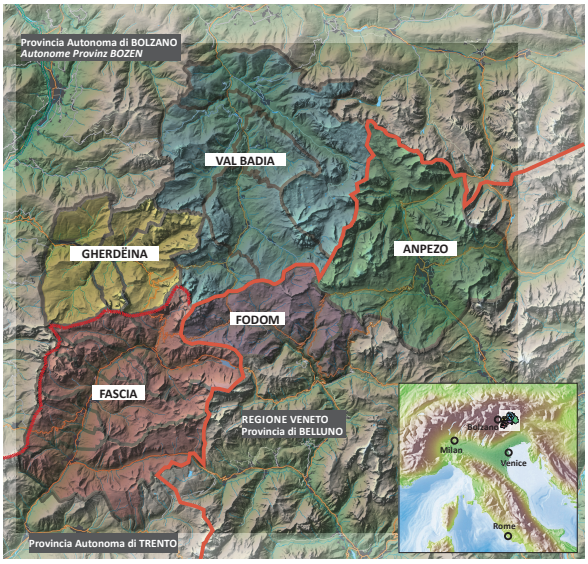

Figure 1: The five Ladin-speaking valleys located across three provinces in northern Italy: Val Badia (blue) and Gherdëina (yellow) in the Autonomous Province of Bolzano/Bozen South Tyrol, Fascia (red) in the Autonomous Province of Trento, Fodom (purple) and Anpezo (green) in the Province of Belluno.

Essentially, developing machine translation systems has actually become easier as it is no longer just about language expertise. Instead, the generalizability capabilities of language models are being exploited, and the limitations now lie more in providing training data and infrastructure. However, its development depends on reliable and extensive test data, which is essential for consistent evaluation, model comparison, and tracking progress over time. The FLORES+ dataset provides such a benchmark and covers more than 200 languages (NLLB Team et al., 2024). It is widely used in research on

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 1061–1071
November 8–9, 2025.

low-resource languages because of its broad language coverage and its ability to support evaluation across more than 19,900 language pairs.

With this in mind, we release the FLORES+ translations for the Ladin language. Ladin is a Rhaeto-Romance language that originated from Vulgar Latin during the Roman conquest of the Alps. It evolved in isolated Alpine valleys, leading to the diverse regional varieties spoken today in Northern Italy (Bauer, 2022). As a result of the fragmented development over centuries there is no single, unified spoken variety for Ladin that is characterised by its internal linguistic diversity, with five main regional variants depicted in Figure 1: *Val Badia*, *Gherdëina*, *Fassa*, *Fodom*, and *Anpezo*. In this work, we extend FLORES+ with translations for Val Badia and Gherdëina. Both variants are spoken in South Tyrol and together representing the largest portion of Ladin speakers. The dataset translation was carried out by professionals at the Ladin Cultural Institute "Micurá de Rü", the primary institution dedicated to preserving and supporting the development of the language.

Moreover, we release parallel datasets for Val Badia–Gherdëina and Gherdëina–Italian that can be used for fine-tuning or retrieval-augmented prompting for this language. These two datasets enables to provide an overview of the performance of `NLLB` and state-of-the-art Large Language Models (LLMs), including `GPT-3.5`, `GPT-4o`, `Llama-3.3`, and `DeepSeek-R1` on Ladin (Val Badia and Gherdëina).

In summary, the contributions of our work are:

1. We submit the FLORES+ translations for Val Badia Ladin (full) and Gherdëina Ladin (`dev` split)

2. We release two parallel datasets of approximately 18,000 sentence pairs for Gherdëina–Italian and Gherdëina–Val Badia.

3. We benchmark state-of-the-art machine translation systems and LLMs on this dataset, identifying current limitations and outlining opportunities for advancing MT for Ladin.

With this, we aim to increase the visibility of Ladin within the MT research community, highlight the ongoing need for focused research on this language, and provide resources that empower researchers and developers worldwide to advance Ladin machine translation.

2 The Ladin Language

Ladin should not be confused with *Ladino* (`lad`), a Judeo-Spanish language. This confusion is common, especially in Italian where Ladino can refer to both; to avoid ambiguity, the term *Ladino delle Dolomiti* (Dolomite Ladin) is often used. Ladin is identified by the ISO 639-3 code `lld`.

For a comprehensive treatment of the language's history, dialectal variation, and sociolinguistic context, we refer the reader to Pescosta (2015); Videsott et al. (2020); Bauer (2022). For a recent contribution discussing the state of the Ladin language in the Dolomite region, we refer to Videsott (2023); Colcuc (2024).

2.1 Historical Development

Ladin is a Romance language that is traditionally associated with Romansh of the Grisons in Switzerland and Friulian of north-eastern Italy within the Rhaeto-Romance group (Videsott et al., 2020). It traces its origins to the Vulgar Latin introduced during the Roman conquest of the Alpine region around 15 BC. Over time, the native populations gradually adopted this Latin, which absorbed elements of pre-Roman languages such as Celtic and Raetic. With the fall of the Western Roman Empire, the evolving Romance dialects began to diverge, influenced by the surrounding Germanic languages and political fragmentation. As Germanic tribes advanced, the speakers of early Ladin were pushed into isolated Alpine valleys, where the language continued to evolve separately from other Romance varieties. This geographic and historical isolation laid the foundation for the internal diversity of Ladin seen today in the Dolomite region of Northern Italy where each variant reflects the unique historical trajectory and degree of contact with neighbouring languages.

The internal linguistic diversity of Ladin has not only historical roots but has also been reinforced by political-administrative fragmentation in the 20th century. Following the annexation of South Tyrol by Italy after the Treaty of Saint-Germain (1919) and the Italianisation policies introduced under the Fascist regime, Ladin-speaking territories were divided among three different provinces: *Val Badia* and *Gherdëina* became part of the autonomous province of Bolzano (South Tyrol), *Fassa* was integrated into the province of Trento, and *Fodom* and *Anpezo* were assigned to the province of Belluno

in the Veneto region (see Figure 1[1]). This fragmentation led to differing levels of legal recognition, educational support, and public visibility for Ladin communities depending on the province.

2.2 Contemporary Situation

The degree of institutional support and practical implementation of Ladin varies among the different regions. In South Tyrol, Ladin enjoys strong official recognition and robust backing in administration, education, and media. According to the 2024 South Tyrol census[2], approximately 4.41% (19.853) of South Tyroleans declared themselves as belonging to the Ladin language group. The highest proportions of Ladin speakers are found in Val Badia, where 97% of the population, and in Val Gardena, where 85% of the population, identify as Ladin-speaking. This corresponds to around 20,000 people Ladin speakers in South Tyrol.

In the Trentino (Fassa Valley), Ladin is also officially recognized but coexists with a stronger presence of Italian, resulting in comparatively less institutional support and daily use. Based on the 2021 census[3], 2.9% (15.775) of residents identified as Ladin speakers in Trentino.

Meanwhile, in the Veneto valleys of Livinallongo and Ampezzo, its status is more limited, and the language faces greater challenges regarding vitality and institutional presence. The precise number of native Ladin speakers in the traditional (formerly Austrian) Dolomites Ladin area of Veneto (namely Fodom and Anpezo valleys) is not available, but can be estimated to be around 4-6,000[4]. Together, these figures amount to an estimated 40.000 Ladin speakers.

Three cultural institutes, set up as public bodies in the three different Provinces, with a total staff of around 20 people[5] work on standardisation, the creation of digital tools, and the promotion of the social status of the language, following the three traditional axes of language policy for minority languages (status-, corpus- and acquisition planning; see Iannaccaro and Dell'Aquila (2004) for further details). Moreover, a Chair of Ladin Language and Culture Studies has been established at the Free University of Bolzano (*Université Lieida de Bulsan*)[6].

2.3 Linguistic and MT Research on Ladin

Linguistic research on Ladin has a long tradition and remains very active, notably through the journal *Ladinia*[7], which has appeared 49 times since its inception in 1977. Another journal devoted to Ladin studies is *Mondo Ladino*[8], also founded in 1977.

While few published parallel datasets exist for Ladin, existing experiments indicate that state-of-the-art machine translation techniques can be adapted to this low-resource language. Frontull and Moser (2024b) developed an NMT system for Val Badia Ladin and compared it to rule-based and statistical systems; Frontull and Moser (2024a) studied back-translation using `GPT-3.5`; and Valer et al. (2024) created a bidirectional system for Fassa Ladin using multilingual training and knowledge transfer, comparing it to `GPT-4o`. However, the test data used in these studies was limited in scope, preventing broader comparability of results and restricting evaluation to Ladin–Italian translation. There is still much potential to explore how to make even better use of the available resources.

2.4 Ladin of Val Badia and Gherdëina

Although ISO 639-3 assigns a single code (`lld`) to Ladin, it may be more accurate to consider `lld` as a macrolanguage encompassing at least five standardized written variants: Ladin of *Val Badia*, *Gherdëina*, *Fascian*, *Fodom*, and *Anpezan*. In this work, we focus on and provide translations for the two written standards of Val Badia Ladin and Gherdëina Ladin, which can be specifically identified using the IETF BCP 47 language tags[9] `lld_valbadia` and `lld_gherd` respectively. These correspond to Glottolog code `gard1241`[10] for Gherdëina. For Val Badia, there is no exact correspondence, as the spelling unifies `badi1244`

[1] An interactive map is available at https://atlantilinguistici.smallcodes.com/ladinia.html

[2] https://astat.provinz.bz.it/de/publikationen/ergebnisse-sprachgruppenzahlung-2024

[3] http://www.statistica.provincia.tn.it/binary/pat_statistica_new/popolazione/Sintesi_Rilevazione_minoranze_2021.1651135663.pdf

[4] a conservative estimate informed by anecdotal evidence, calculated for a total population of 7,000

[5] a comparatively large number given the size of the population

[6] https://www.unibz.it/

[7] https://www.micura.it/de/ativites-2/ladinia, ISSN 1124-1004

[8] https://www.istladin.net/en/pubblicazioni, ISSN 1121-1121

[9] https://www.iana.org/assignments/language-subtag-registry

[10] https://glottolog.org/resource/languoid/id/gard1241

English	The walls and roofs of ice caves can collapse and cracks can get closed.
Italian	Le pareti e il soffitto delle caverne di ghiaccio sono soggetti a crolli e le crepe possono richiudersi.
Val Badia	I parëis y le sössot di andri da dlacia pó tomé ite y les sfësses pó se stlüje pro.
Gherdëina	I parëies y i plafons dla ciavernes tla dlacia possa tumé ite y la sfëntes possa se stlù.

Table 1: Example translations from the FLORES+ dev split.

and mare1258, with badi1244[11] being the more appropriate mapping.

Linguistic Features Socio-linguistically, in Gardena and Badia valleys Ladin is in contact both with German (in the local dialectal form and in the standard official form) and Italian, in the remaining valleys Italian is the dominant contact language (Videsott et al., 2020). Being Ladin a romance language, standard Italian is in any case a natural point of comparison to highlight important linguistic differences that affect machine translation. Ladin differs from Italian in several ways that are relevant for MT. In the following, we summarize key differences taken from (Bauer, 2022) Phonologically, Ladin features voicing of intervocalic stops (e.g., Latin *p*, *t*, *k* become voiced between vowels), retention of final *-s* to mark plurals, and palatalisation of *c*, *g* before *a*. These sound changes are consistently reflected in Ladin spelling, which means the written forms differ systematically from Italian. Morphologically, Ladin has a complex pattern of plural endings (*-s*, *-i*, other form of palatised consontant) that Italian lacks. Additionally, Ladin has a high number of loanwords from Germanic languages due to long-standing contact (especially for northern varieties), with southern varieties borrowing mainly from Italian, resulting in significant lexical variation even within Ladin written and oral varieties. These phonological, morphological, and lexical differences shape the vocabulary and structure that MT systems must handle, making their accurate representation essential for building effective Ladin machine translation models and a compelling case for natural language processing research.

To give an intution on the similarity between the two variants and Italian and on the difficulty of the translation task, we computed the BLEU score obtained by leaving the text untranslated, which resulted in a score of 5.0 for Italian–Val Badia, 4.3 for Italian–Gherdëina and 12.9 for Val Badia–Gherdëina. Table 1 presents a sample translation of the English sentence *"The walls and roofs of ice caves can collapse and cracks can get closed."* into Val Badia and Gherdëina, as found in the submitted dev split of FLORES+, illustrating these similarities.

Relevant Resources Although relatively little previous research has focused on machine translation for Ladin, significant and valuable work has been carried out in the development of the language itself. Comprehensive dictionaries have been compiled, alongside essential language reference materials, including books detailing grammar and spelling rules. In the following we list relevant resources for the variants in focus in this work developed by the Ladin Cultural Institute "Micurá de Rü":

- *Grafia nöia - Ladin scrit dla Val Badia* (Mischí et al., 2015) is a practical guide to the standardized orthography of Ladin as used in Val Badia which was reformed in 2015.
- *La ortografia dl ladin de Gherdëina* (Forni, 2019b) is a practical guide to the standardized orthography of Ladin as used in Val Gardena.
- *Dizionario italiano - ladino Val Badia* (Moling et al., 2016) is a bilingual dictionary published in 2016. The dictionary includes 30,829 Italian lemmas and 32,701 Ladin lemmas, along with 18,120 phraseological expressions. It offers morphological and encyclopedic information.
- *Dizionario italiano - ladino gardenese* (Forni, 2013) is a bilingual dictionary published in 2013. Authored by Marco Forni, this work is an essential resource for the Ladin language as spoken in Val Gardena (Gherdëina). This extensive bilingual dictionary, contains over 67,000 entries and nearly 20,000 phraseological expressions, provides detailed lexical information along with contextual examples.
- *Gramatica Ladin Gherdëina* (Forni, 2019a), authored by Marco Forni and published in

[11] https://glottolog.org/resource/languoid/id/badi1244

2019. Covering phonetics, morphology, and syntax, it is a comprehensive grammatical reference for the Ladin language as spoken in Val Gardena.

- Several *technical and domain-specific glossaries* are available for Ladin, including, for example, mobile phone interfaces fully translated in collaboration with Motorola (Oliveira et al., 2024) and terminology in pedagogy[12]. Glossaries in other domains, such as music, animals and plants, history, and historiography, are currently under elaboration.

- *Spellchecker*: online tools[13] as well as Firefox Add-ons are available for Val Badia[14] and Gherdëina. [15]

- set of *18k parallel sentences for Ladin Val Badia–Italian*, as used in (Frontull and Moser, 2024a), available at HuggingFace[16]

We leveraged the existing dictionaries to implement spelling correction and extracted the parallel sentences contained in the datasets we release these resources.

3 Data collection

The translation of the FLORES+ dataset into Ladin was carried out in close collaboration with the Ladin Cultural Institute "Micurá de Rü"[17]. In this section, we provide an overview of the translators involved, the tools used, the procedures followed for the different Ladin variants, and the quality assurance measures implemented.

Translators The translation team consisted of five translators for Val Badia and two translators for Gherdëina. All translators involved in this process are (or were at the time of the project) employed by the Ladin Cultural Institute, where they work professionally with Ladin and therefore have extensive experience using the language in both spoken and written forms. They are native speakers of Ladin and hold a C1-level certification in Ladin, obtained through the official language exams[18] regulated by the Autonomous Province of Bolzano/Bozen, making them highly valuable collaborators, whose contributions ensure reliable, high-quality work. All translators also have a good understanding of English (though not formally certified) and are fluent at C1 level in both German and Italian. Prior to beginning the task, all translators were asked to carefully read and acknowledge the OLDI translation guidelines[19], which explicitly require human translation, prohibit the use of machine translation systems and provide detailed instructions on tone, consistency, fidelity to the source, and the handling of named entities. In accordance with these guidelines, translators were instructed to base their work primarily on the English source texts. However, due to their proficiency in German and Italian, these translations were also provided as additional references. Due to the linguistic similarities between Ladin and Friulian, these texts were also included as reference texts.

Translation Assignment Tool The work was divided into kits of 25 sentences. To manage and assign the translation work, a dedicated web-based platform was set up. This tool allowed kits to be assigned to specific translators and the progress of the translation to be monitored, eliminating the need to distribute and manage separate text files. The tool displays the original English text along with reference translations into German, Italian, Friulian, and, for Gherdëina, also the translation into Val Badia. Translators can enter their Ladin translation directly into an input field. The tool also has a built-in spell checker that highlights unknown or potentially misspelled words and helps users identify and correct typos. Figure 2 shows a screenshot of the user interface for a sentence to be translated, with the English source text, reference translations, and the input field for the Ladin translation. Sentences can also be skipped and revisited later.

Val Badia Translations In a first phase, the sentences were translated into the Ladin Val Badia. Each translator received two kits per week, a work-

[12] https://pedagogia.ladinternet.it/glossary

[13] https://www.micura.it/en/online-services/spellchecker

[14] https://addons.thunderbird.net/de/thunderbird/addon/lld-valbadia/

[15] https://addons.thunderbird.net/de/thunderbird/addon/lld-gherd/

[16] https://huggingface.co/datasets/sfrontull/lld_valbadia-ita

[17] https://micura.it

[18] The exams include listening, writing, speaking, and reading comprehension and are offered in both the Gherdëina and Val Badia varieties. This certification is a prerequisite for a permanent employment at the Ladin Cultural Institute in South Tyrol. For more information we refer to https://zweisprachigkeitspruefungen.provinz.bz.it/de/ladinischpruefung

[19] https://oldi.org/guidelines

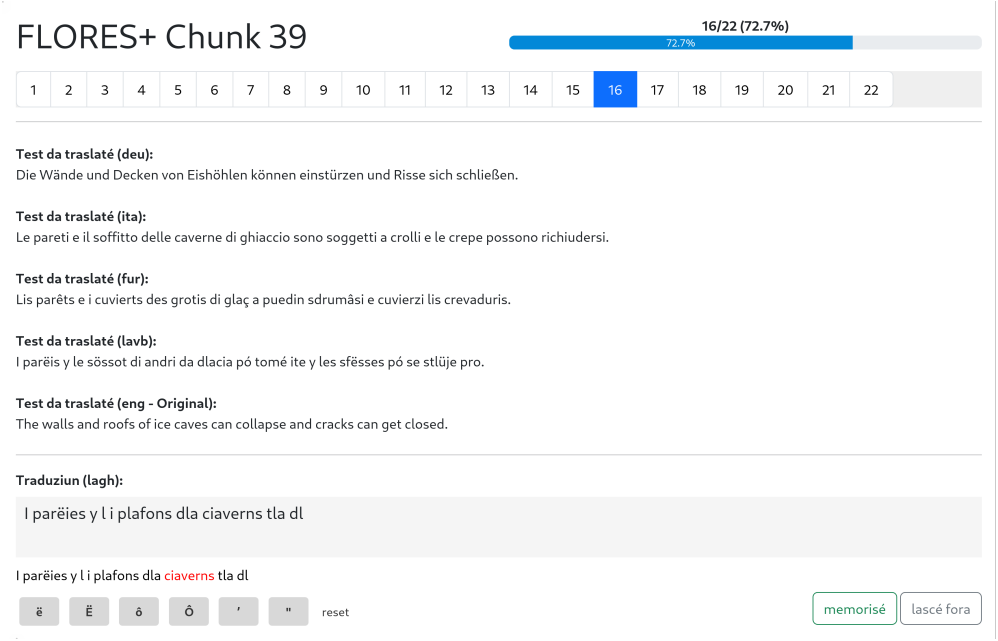

Figure 2: Screenshot of the tool developed for the translation of FLORES+.

load that proved manageable alongside daily tasks. This modular approach ensured an even distribution of work, allowed careful review of each unit, and maintained a sustainable pace. Overall, the team was able to complete the work within four months.

Gherdëina Translations Building on the Val Badia translations, we began translating into Ladin Gherdëina, using the Val Badia translations as an additional reference. This sequential approach was practical, as adapting an existing Ladin text is more efficient than translating from scratch, particularly with limited personnel. This second translation cycle also served as an implicit revision of the Val Badia texts, during which minor errors were identified and corrected. Given the smaller number of translators, the project timeline has been adjusted to ensure careful and thorough completion (work is still in progress).

Quality Assurance The first author, also a native Ladin speaker, continuously reviewed a sample of the generated translations and extracted all unknown words detected with a variant-specific spell checker. These were compiled into a shared document, where translators were asked to review and confirm or correct the entries. Words requiring concordance across variants were fed back to the group, ensuring consistency. This collaborative monitoring helped to identify and resolve potential errors. Concerns about the use of external machine translation systems (e.g., Google Translate or ChatGPT) do not apply in this setting. Major MT systems do not support Ladin, nor its specific variants, and ChatGPT performs poorly on Ladin (as evidenced in our evaluation results below). Any attempt to use such tools would have been immediately detectable, as they are incapable of generating variant-specific vocabulary and would trigger numerous alerts in the spell checker.

Neologisms The translation activity was accompanied by intensive collaboration among the translators. This collaborative process not only helped resolve stylistic and terminological questions but also led to the creation of new words that had not previously existed in the respective variety. As such, the project contributed to the lexical development of the Ladin language. In total, around 500 new words were created. Examples of such words are: *codificades* (encoded), *dejabilité* (disability), *surafolamënt* (overcrowding) or *triceratop*.

4 Experimental Validation

In this section, we provide an overview of the results achieved with a NMT system and different state-of-the-art LLMs on the manually created FLORES+ translations. The aim is to assess the current capabilities of these models in translating between Val Badia, Gherdëina and Italian, both in off-the-shelf zero-shot settings and using the accompanying parallel data as fine-tuning training

data for the NMT system and as retrieval data for the LLMs. Moreover, this serves as validation of the quality of the submitted datasets.

Training and Retrieval Corpora For Italian–Val Badia, a dataset is already available[20] consisting of word usage example sentences extracted from the Italiano–Ladin Val Badia dictionary (Moling et al., 2016). To create training and retrieval resources for Gherdëina, we extracted the word usage example sentences contained in the Italiano–Ladin Gherdëina dictionary (Forni, 2013). Since the two dictionaries are based on a common foundation and were largely coordinated with each other, they contain many common example sentences. This overlap allowed us to construct a Val Badia–Gherdëina parallel dataset by aligning identical Italian sentences. This yielded the following datasets for training (for `NLLB`) and retrieval (for LLMs):

- 18,140 sentences for Val Badia–Italian,
- 19,971 sentences for Gherdëina–Italian, and
- 14,953 sentences for Val Badia–Gherdëina.

Note that, given their purpose of illustrating word usage, the sentences in these corpora are relatively short and simple, with an average length of approximately 25 characters. We make both the Gherdëina–Italian[21] and Val Badia–Gherdëina[22] datasets publicly available under a `CC BY-NC-SA 4.0`[23] license.

4.1 Neural Machine Translation Models

For this evaluation, we tested the performance of the multilingual NMT system `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2024). We evaluated this system off-the-shelf (without any fine-tuning) from Ladin to Italian and after being fine-tuned with the parallel datasets available for Val Badia–Italian, Gherdëina–Italian and Val Badia–Gherdëina mentioned above. We trained a single model covering all translation directions simultaneously using the `transformers` framework for 10 epochs with a batch size of 16 (which corresponded to approximately 101,930 optimization steps) and a maximum sequence length of 196 tokens. The validation loss steadily improved throughout training. Training was completed in 7 hours and 42 minutes on an NVIDIA RTX 4090 GPU.

4.2 Large Language Models

LLMs are proving to be increasingly valuable translators. A key advantage of LLMs is their flexibility in adapting to different specifications and contexts, which enables more targeted, application-specific use. Several techniques have also been studied for machine translation in low-resource languages, showing that LLMs can perform well in these settings and can often be further improved by providing additional data (Agrawal et al., 2023; Tang et al., 2024). Motivated by this, we evaluated different LLMs on Ladin to gain an overview of their performance and the results they can achieve. We evaluate the following four LLMs: (*i*) `GPT-3.5` is a general-purpose language model from OpenAI's GPT-3 series with 175B parameters, released in 2022 (Brown et al., 2020). (*ii*) `GPT-4o` a model by OpenAI, introduced in 2023, with 200B parameters and enhanced reasoning capabilities designed for complex problem-solving (Hurst et al., 2024). (*iii*) `Llama-3.3` is a text-only model by Meta AI, released in December 2024, featuring 70B parameters (Touvron et al., 2023). (*iv*) `DeepSeek-R1` is a reasoning-focused model by DeepSeek AI, introduced in January 2025, with 658B parameters (DeepSeek-AI et al., 2025). The models were prompted using the API services: OpenAI API[24] for `GPT-3.5` and `GPT-4o`, DeepSeek API[25] for `DeepSeek-R1`, and Together Inference API[26] for `Llama-3.3`. The hyperparameters were configured according to the default settings provided by each service.

Due to the very limited amount of machine-readable data available for Ladin it is likely that LLMs have minimal exposure to Ladin and are unaware of its internal variation.

4.2.1 Prompting Techniques

We applied two prompting methodologies in our experiments: *zero-shot* and *BM25-Retrieval*. The *zero-shot* method (Robinson et al., 2023; Gao et al., 2024; Hendy et al., 2023; Bawden and Yvon, 2023) relies solely on the model's pre-existing knowledge. The prompt directly instructs the model to translate sentence into the target, without providing explicit

[20] https://doi.org/10.57967/hf/1878

[21] https://huggingface.co/datasets/sfrontull/lld_gherd-ita

[22] https://huggingface.co/datasets/sfrontull/lld_valbadia-lld_gherd

[23] https://creativecommons.org/licenses/by-nc-sa/4.0

[24] https://platform.openai.com

[25] https://www.deepseek.ai/api

[26] https://www.together.ai

Model	low- to high-resource VB→IT	low- to high-resource GH→IT	high- to low-resource IT→VB	high- to low-resource IT→GH	low- to low-resource VB→GH	low- to low-resource GH→VB
zero-shot / base model						
GPT-3.5	19.35±2.09	18.52±2.04	4.91±1.23	6.73±1.18	10.52±1.41	10.73±1.53
GPT-4o	**22.90**±2.11	**23.03**±2.09	5.70±1.40	5.53±1.20	11.15±1.61	11.17±1.37
Llama-3.3	20.31±2.10	21.02±2.12	6.46±1.29	9.50±1.35	15.01±1.69	12.46±1.53
DeepSeek-R1	22.47±2.04	23.02±1.96	6.31±1.24	8.77±1.30	13.11±1.52	10.19±1.34
NLLB BM	14.86	12.49	-	-	-	-
BM25-Retrieval / fine-tuned						
GPT-3.5	26.68±2.03	20.95±1.69	9.64±1.27	8.22±1.19	15.95±1.49	16.00±1.68
GPT-4o	**29.47**±2.25	**23.51**±2.16	14.49±1.55	11.62±1.41	22.72±1.97	19.58±2.00
Llama-3.3	27.20±2.04	22.33±1.98	15.50±1.77	13.20±1.51	24.35±1.89	21.22±1.98
DeepSeek-R1	28.91±1.93	22.30±1.83	16.06±1.66	14.07±1.46	24.81±1.78	22.49±2.08
NLLB FT	21.73	17.80	**18.06**	**14.48**	**29.63**	**31.15**

Table 2: BLEU mean scores and confidence intervals for each model across six translation directions, split by method (*zero-shot* and *BM25-Retrieval / fine-tuned*).

Model	IT→VB	IT→GH	VB→GH	GH→VB
reference	78%	80%	80%	78%
GPT-3.5	53%	52%	55%	58%
GPT-4o	58%	58%	64%	65%
Llama-3.3	60%	62%	68%	66%
DeepSeek-R1	63%	64%	67%	68%
NLLB FT	**76%**	**73%**	**76%**	**79%**

Table 3: Average proportion of words in manually created ladin translations and those generated with BM25-Retrieval/fine-tuned NLLB model passing spellcheck.

translation examples or lexical guidance. This baseline approach tests the model's intrinsic understanding of Ladin syntax and vocabulary. To enhance translation quality beyond zero-shot capabilities, we employ *BM25-Retrieval* to provide the model with relevant in-context examples based on lexical similarity. BM25 (Robertson et al., 1995) is a probabilistic ranking method that estimates the relevance of documents to a given search query. This sparse retrieval method ranks examples based on lexical-level overlap with the input sentence, aiming to offer relevant in-context examples grounded in lexical similarity. It has been shown to be highly effective also for retrieving examples for MT with LLMs (Agrawal et al., 2023; Tang et al., 2024). We implemented this method using the `bm25s` Python package[27] (Lù, 2024) to select 30 translation examples from the corresponding retrieval corpus that we included in the prompt.

5 Results

Table 2 presents the BLEU scores for the six translation directions between Val Badia (VB), Gherdëina (GH), and Italian (IT). The upper half of the table reports results obtained with the models "off-the-shelf," that is, using zero-shot prompting for `GPT-3.5`, `GPT-4o`, `Llama-3.3`, and `DeepSeek-R1`, as well as the NLLB base model (BM). The lower half shows the results obtained when exploiting additional parallel data as retrieval data and to fine-tune the NLLB model (FT). Since the LLMs were evaluated on a subset of 175 sentences, we report the average BLEU (Post, 2018) scores together with the standard deviation[28] to provide a confidence interval for the results. The best score in each translation direction, when comparing the LLMs with NLLB, is highlighted in bold. In Table 3, we report the average portion of words in the generated translations that pass spellcheck. This should give an idea of the general lexical quality of the translations.[29] Also in this table we highlighted the best scores (LLMs vs. fine-tuned NLLB model) in bold.

[27] https://github.com/xhluca/bm25s

[28] Computed with `sacrebleu` (Post, 2018)

[29] Note that the reference translations do not achieve 100% spellcheck accuracy primarily because many named entities are absent from the dictionary and thus flagged as invalid. Therefore, this metric serves mainly as an expectation indicator rather than an absolute measure of correctness.

Interestingly, concerning the fine-tuned NLLB model, the relatively simple training datasets proved to be effective, especially considering that the benchmark data is substantially more complex. Despite the brevity and simplicity of the training samples, the model generated complete translations of the test samples. One peculiarity we observed is that the generated translations usually start with lowercase letters, reflecting the format of the training examples. Another limitation is that the model struggles with morphological variations: since most training examples contained words in their infinitive form (e.g., "I go home" and never "he went home"), the model is unable to generate correct inflections or conjugations. LLMs are less influenced by the simple translation examples and do not "compromise" their fluency when translating into Italian. For example, they also use the correct capitalization at the beginning of the sentence. When translating from Italian into Ladin, however, LLMs often lack the appropriate vocabulary, whereas the fine-tuned NLLB model adapts the vocabulary more effectively (see Table 3). In contrast, translation between Ladin variants is an easier task, as it mainly involves adapting the vocabulary. Here, the NLLB model already delivers satisfactory results and the difference between NLLB and LLMs is more pronounced, highlighting the challenge LLMs face in accurately adapting vocabulary.

From these results, three main observations can be drawn: (i) current LLMs exhibit limited coverage of Ladin variants, with translation into Ladin remaining a clear challenge; (ii) incorporating the parallel data released in this work yields substantial improvements in translation quality across models, but limitations remain in fluency and morphological variation due to the simplicity of the training examples; (iii) the relative advantage of the systems depends on the translation direction: when translating from a low-resource to a high-resource language, LLMs enhanced with retrieval-augmented generation achieve the best results, whereas for high-to-low-resource translation, the fine-tuned `NLLB` model performs better. This conclusion is further supported by a significantly higher proportion of valid words in its generated translations. The incorporation additional data (e.g. through back-translation) would yield better results; however, the primary focus here is on validating the quality of the provided datasets. Overall, these findings underscore both the potential and the necessity of advancing machine translation research for Ladin, as well as the value of the datasets we contribute.

6 Conclusion

In this work, we present our submission to the OLDI shared task 2025, providing FLORES+ translations for Ladin (Val Badia and Gherdëina) and provide an evaluation of several LLMs and the `NLLB` model on this test set, covering six translation directions between Val Badia, Gherdëina, and Italian.

Our results show significant performance gains from the additional parallel datasets released with this work, validating the quality of the datasets and highlighting a promising direction for future research. Beyond fine-tuning neural MT models or relying on basic BM25-Retrieval, our datasets opens the door to more advanced retrieval-augmented prompting strategies, where semantically or syntactically similar sentence pairs are selected to guide translations (Kumar et al., 2023; Merx et al., 2024; Tang et al., 2024; Zebaze et al., 2025).

A central question going forward is whether such carefully designed retrieval and prompting methods could not only provide a more lightweight alternative to fine-tuning for low-resource languages like Ladin, but in some cases even surpass it in translation quality. It would be highly valuable if, in future work, this effort could be extended to the Anpezo, Fassa, and Fodom Ladin variants, enabling their inclusion in the FLORES+ dataset and thus fully representing the Ladin language and accurately reflecting its internal diversity. We hope this work will inspire and encourage further research on Ladin in machine translation, helping to bring the language into sharper focus within the MT community.

Acknowledgements

We would like to thank the anonymous reviewers for their work and valuable comments and suggestions, which have greatly helped to improve our presentation. This initiative was taken as part of the research project *Machine Translation for Ladin* at the University of Innsbruck carried out in collaboration with the Ladin Cultural Institute "Micurá de Rü" and funded by the *Regione Autonoma Trentino–Alto Adige*. We also thank Jürgen Runggaldier for his support in this endeavour.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context Examples Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Roland Bauer. 2022. Language. In Tobia Moroder, editor, *The Ladins of the Dolomites*, pages 28–35. PLUS – University of Salzburg, Salzburg and Bolzano.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Beatrice Colcuc. 2024. Il ladino. *Linguistik Online*, 130(6):9–30. Creative Commons Attribution 4.0 International License.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and et. al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Marco Forni. 2013. *Dizionario italiano – ladino gardenese / Dizioner ladino gardenese – italiano*. Istitut Ladin Micurá de Rü, San Martin de Tor.

Marco Forni. 2019a. *Gramatica Ladin Gherdëina*. Istitut Ladin Micurá de Rü.

Marco Forni. 2019b. *La ortografia dl ladin de Gherdëina*. Istitut Ladin Micurá de Rü, San Martin de Tor.

Samuel Frontull and Georg Moser. 2024a. 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.

Samuel Frontull and Georg Moser. 2024b. Traduzione automatica "neurale" per il ladino della Val Badia. *Ladinia*, XLVIII:119–144.

Yuan Gao, Ruili Wang, and Feng Hou. 2024. How to Design Translation Prompts for ChatGPT: An Empirical Study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops, New York, NY, USA. Association for Computing Machinery.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gabriele Iannaccaro and Vittorio Dell'Aquila. 2004. *La pianificazione linguistica: lingue, società e istituzioni*. Carocci, Roma.

Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining Multiple Features for In-context Example Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Giovanni Mischí, Claudia Rubatscher, Isabella Ties, Daria Valentin, and Paul Videsott. 2015. *Grafia nöia - Ladin scrit dla Val Badia: por les scolines y les*

scores ladines. Departimënt Educaziun y Cultura ladina, Balsan.

Sara Moling, Ulrike Frenademetz, and Marlies Valentin. 2016. *Dizionario Italiano–Ladino Val Badia/Dizionar Ladin Val Badia–Talian*. Istitut Ladin Micurá de Rü, San Martin de Tor.

NLLB Team, 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, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Janine Oliveira, Marison Ranieri Rodrigues de Freitas, Delaney Gomez-Jackson, Juliana Peres Rebelatto Pereira, Natalia Sarmento Tenório Falcão, Roy Yokoyama, Sushil Garg, and Yukitomi Fujinaga. 2024. *Hello Indigenous: a blueprint on the preservation of endangered Indigenous languages through digital inclusion*. Lenovo Foundation and Motorola Mobility, Brazil. UNESCO-sponsored; Electronic version.

Werner Pescosta. 2015. *Storia dei ladini delle Dolomiti*, 2a edizione rielaborata edition. Istitut Ladin Micurà de Rü, San Martin de Tor, Italy.

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, Belgium, Brussels. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Chenming Tang, Zhixiang Wang, and Yunfang Wu. 2024. SCOI: Syntax-augmented Coverage-based In-context Example Selection for Machine Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9956–9971, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Giovanni Valer, Nicolò Penzo, and Jacopo Staiano. 2024. Nesciun Lengaz Lascià Endò: Machine Translation for Fassa Ladin. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 967–975, Pisa, Italy. CEUR Workshop Proceedings.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need.

Paul Videsott. 2023. *Les Ladins des Dolomites*. Peuples en Péril. Éditions Armeline, Crozon.

Paul Videsott, Ruth Videsott, and Jan Casalicchio, editors. 2020. *Manuale di linguistica ladina*, volume 26 of *Manuals of Romance Linguistics*. De Gruyter.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.