Towards the First NLP Benchmark for Ladin - an Extremely Low-Resource Language

Ulin Nuha*
Toyota Technological Institute, Japan
ulin_nuha@toyota-ti.ac.jp

Adam Jatowt
University of Innsbruck, Austria
adam.jatowt@uibk.ac.at

Abstract

The performance of large language models (LLMs) tends to degrade for extremely low-resource languages, primarily due to the lack of labeled training data. Despite growing interest, the availability of high-quality natural language processing (NLP) datasets for these languages remains limited. This paper addresses such gap by focusing on Ladin, an endangered Romance language, specifically the Val Badia variant. Leveraging a small set of parallel Ladin–Italian sentence pairs, we create synthetic datasets for sentiment analysis and question answering by translating monolingual Italian data. To ensure linguistic quality, we apply rigorous filtering and back-translation procedures in our method. We further demonstrate that incorporating these synthetic datasets into machine translation training leads to substantial improvements over existing Italian–Ladin translation baselines. Our contributions include sentiment analysis and question answering datasets for Ladin, establishing foundational resources that support broader NLP research and downstream applications for underrepresented languages.

1 Introduction

Large language models (LLMs) have garnered significant attention in diverse audiences (Wang et al., 2024; Gambardella et al., 2024) due to their ability to effectively perform natural language tasks with only a few input-output examples (Cheng et al., 2024), while also eliminating the need for model's gradient updates (Nguyen et al., 2024). LLMs achieve remarkable performance through pre-training on large corpora, but their reliance on high-resource languages (HRLs) limits their effectiveness for low-resource languages (LRLs) (Pham et al., 2024), especially extreme ones (Purason et al., 2024). Efforts to address this limitation include various techniques such as in-context learning (Cahyawijaya et al., 2024) and fine-tuning (Alabi et al., 2022; Su et al., 2024) to transfer LLM capabilities to LRLs. Additionally, some studies (Yong et al., 2024; Morim da Silva et al., 2024; Tran et al., 2024) have focused on developing translation systems between LRLs and HRLs to improve their translation performance. The limited advancement of natural language processing (NLP) in LRLs stems largely from the disproportionate focus on HRLs, often at the expense of tools and resources for low-resource languages (Zhang et al., 2024). Moreover, the scarcity of NLP datasets for extremely low-resource languages (ELRLs) remains a major barrier, highlighting the need for inclusive AI to support marginalized languages.

Figure 1: The Ladin speaking region in Northern Italy.

This paper advances NLP research for Ladin, an extremely low-resource language in South Tyrol, Northern Italy (see Figure 1). Despite progress in NLP for LRLs, Ladin's dialectal diversity and limited digitized corpora pose challenges. Building on the prior work (Frontull and Moser, 2024), which established a machine translation (MT) benchmark

*Work done while the author was a PhD student at National Kaohsiung University of Science and Technology, Taiwan.

Findings of the Association for Computational Linguistics: EACL 2026, pages 1049–1064

between Italian and Ladin (Val Badia variant), we further improve translation performance for Ladin.

Ladin is a Rhaeto-Romance language with around 30,000 speakers. It comprises five dialects (Val Badia, Fascia, Anpezo, Fodom, and Gherdëina) which vary significantly in morpho-syntactic and orthographic conventions. Although standardization efforts like Ladin Dolomitan exist, it is not widely recognized and is used sparingly, with speakers reporting limited familiarity beyond their own varieties (Connor, 2023). Linguistic conventions have also changed over time within dialects. This linguistic diversity presents major challenges for machine translation and resource creation. Publicly available resources are extremely limited, with only a few lexicons, dictionaries, and minor corpora. A major source of monolingual Ladin data is *La Usc di Ladins*[1], a newspaper published in five variants and digitally archived, but this resource remains underutilized due to lack of alignment and limited annotation. Our work focuses on the Val Badia variant, due to the largest accessible dataset. Previous work on MT for Ladin's Fascia variant exists (Valer et al., 2024); however, the parallel dataset used contains only 1,135 sentences.

Unlike prior studies, our work extends beyond MT for Ladin to include sentiment analysis (SA) and multiple-choice question answering (MCQA), establishing the first Ladin datasets for these tasks. First, we compare MT approaches, including few-shot learning and fine-tuning, using available paired Italian–Ladin data. Then, we construct high-quality Ladin datasets for SA and MCQA through the translation from monolingual Italian data, applying rigorous filtering to ensure high quality. These datasets not only contribute to SA and MCQA but also enhance MT. We also compare our synthetic dataset with existing Italian–Ladin translation benchmarks and evaluate its impact.

In summary, the primary contributions of our work can be outlined as follows:

- We conduct a comparative study on translation between Italian and Ladin (Val Badia variant). Our model significantly outperforms the current benchmark, demonstrating the effectiveness of our approach for this extremely low-resource language (LRL).
- We construct a high-quality synthetic Ladin–Italian paired dataset, derived from monolingual Italian data.
- We demonstrate the utility of the synthetic dataset beyond MT by applying it to downstream NLP tasks such as sentiment analysis and question answering, establishing the first such datasets in Ladin.

[1]https://www.lausc.it/

2 Background and Related Work

NLP for ELRLs has garnered attention from researchers due to the enduring challenges with data scarcity. These languages typically possess fewer than 0.1 million available parallel sentence pairs (Ranathunga et al., 2023), which are insufficient for effectively training neural machine translation (NMT) models (Murthy et al., 2019). Notably, Ladin has fewer than 100 thousand parallel sentences available for MT (Frontull and Moser, 2024), while other NLP resources, such as text classification and question answering, remain non-existent.

Ladin and Italian are Romance languages with shared Latin roots, resulting in potential similarities in lexical and syntactic aspects. This linguistic proximity facilitates cross-lingual transfer, allowing multilingual models trained on Italian to generalize well to Ladin. However, key divergences remain due to Ladin's unique phonological and morphological traits (Melchior, 2023). These differences stem from geographic isolation and varying contact with surrounding languages, particularly regional varieties of German and Italian. Most Ladin speakers in Italy are bi- or trilingual, using Ladin in private domains and Italian, German, or both in public settings, as neighboring communities generally lack comprehension of Ladin (Erardi et al., 2022). This contact has led to regional variation. Gherdëina and Val Badia exhibit stronger German influence, while southern valleys show more Italian features. Val Badia is often perceived as the "purest" Ladin, though this reflects relative rather than absolute isolation from external influence. These dynamics present challenges for translation. The absence of a standard variety necessitates dialect-specific translation strategies, hindering mutual intelligibility and the development of shared resources. Long-standing language contact has introduced borrowings and calques, complicating the identification of core Ladin structures.

To address resource scarcity in ELRLs, recent efforts have leveraged transfer learning, particularly via LLMs (Pham et al., 2024; Lim et al., 2024).

However, success remains limited in such settings (Tran et al., 2024), where extreme data sparsity poses challenges for effective domain adaptation and language alignment. In addition, a key aspect of transfer learning with LLMs involves leveraging token similarity and cross-region similarity to better capture shared cultural and linguistic features (Bagheri Nezhad et al., 2025). Shu et al. (2024) proposed a MT framework that integrates Retrieval-Augmented Generation (RAG) with LLMs to address these challenges, whereas Lu et al. (2025) explored LLMs combined with direct preference optimization to enhance translation quality. Despite promising results, these methods incur high computational costs.

Multimodal approaches of MT, which incorporate modalities such as visual context (Rajpoot et al., 2024; Ul Haq et al., 2024; Hatami et al., 2024), require supplementary datasets that are often unavailable and face considerable computational and complexity challenges. Then, rule-based machine translation (RBMT), exemplified by Apertium (García, 2024; Sánchez-Martínez et al., 2024), offers a less resource-intensive alternative with lower computational cost than NMT. However, RBMT is labor-intensive, requiring extensive manual effort to create and maintain linguistic rules, with scalability further hindered by the challenge of ensuring rule self-consistency (Liu et al., 2023).

Another approach to supporting technologies for LRLs is benchmark development. For example, Urbizu et al. (2022) introduced BasqueGLUE, a benchmark of NLP tasks for Basque. Mokhtarabadi et al. (2025) presented FarsInstruct, a large dataset to strengthen LLMs' capacity to follow instructions in Persian, a globally low-represented language. Moreover, recent work in language revitalization and human-centered NLP shows the need to go beyond performance metrics in developing technologies for endangered languages (Bird, 2020). Effective systems should align with ethical principles and community-driven goals, including intergenerational transmission and cultural continuity.

Following this perspective, our work presents a comparative study of MT approaches, including instruction learning and fine-tuning (Zhang et al., 2023), using language models such as NLLB and GPT-4. To evaluate translation quality, we use both statistical and qualitative analyses, particularly in the context of synthetic data. While prior research has made progress in NLP for LRLs, substantial gaps remain for Ladin. This study addresses these gaps by leveraging data augmentation and introducing benchmark datasets for MT, SA, and MCQA.

3 Experiments

3.1 Data resources

In this section, we provide an overview of data resources used in our study. We construct novel synthetic Ladin datasets for three tasks: MT, SA, and MCQA. Data resources were collected from publicly available datasets.

Machine translation. The paired Italian–Ladin datasets used for training and testing baseline MT models are derived from prior work (Frontull and Moser, 2024). The original Italian-Ladin dataset AD_{Ita_Lad} for training comprises 18,139 sentence pairs, crafted as concise examples of specific word and phrase usage. The average sentence length in AD_{Ita_Lad} is 23.43 characters for Ladin and 25.69 characters for Italian, while the average number of words per sentence is 5.02 for Ladin and 4.36 for Italian. Subsequently, testing dataset is divided into three subsets: Legal (L), 424 sentences, focused on formal and legal terminology; Historical-Legal (H), 833 sentences, combining historical narratives with legal and administrative language; and Literary (Lt), 1,563 sentences, featuring narrative prose, dialogue, and idiomatic expressions. For the Italian portion of the testing dataset, the average word counts in the L, H, and Lt subsets are 21.25, 26.65, and 15.61 words, respectively. In comparison, the corresponding averages in Ladin are 23.58, 24.52, and 13.05 words.

Sentiment analysis. We construct the Ladin datasets by leveraging labeled monolingual Italian resources. For the SA task, we use the monolingual Italian SA dataset D_{Ita_SA} from Desole (2020), which contains 41,077 Tripadvisor reviews labeled as positive or negative. To mitigate the impact of excessively long reviews, we retain only entries with word counts below the third quartile (138 words), reducing the dataset to 30,712 entries. Since the raw dataset collected via web scraping contains grammatical errors, we use GPT-4 to correct them while preserving the original semantics, using few-shot learning shown in Appendix C. After correction, the average sentence length in D_{Ita_SA} is 70.44 words.

MCQA. For the MCQA task, we use the monolingual Italian MCQA dataset D_{Ita_MCQA} from Rinaldi et al. (2024). D_{Ita_MCQA} comprises over 5,000 manually crafted questions covering various topics.

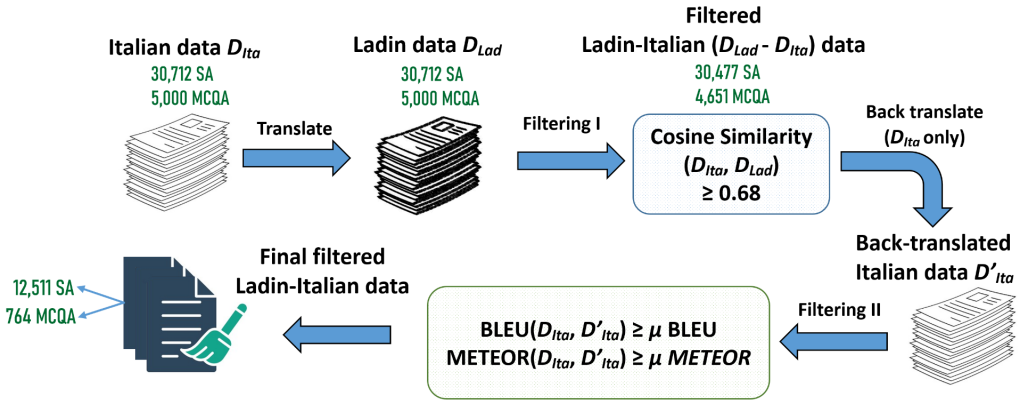

Figure 2: Synthetic data generation process. Initially, we translate the Italian data D_{Ita} of SA and MCQA into Ladin. The Ladin translation D_{Lad} of SA and MCQA is then filtered using Filtering I. Next, the filtered D_{Lad} is back-translated into Italian D'_{Ita}, followed by Filtering II to obtain the high-quality parallel paired datasets.

This MCQA dataset includes questions with 2 to 6 answer choices. However, questions with 2 or 6 options are excluded due to their limited frequency.

3.2 Language model settings for MT

We describe the process of building an MT model using existing Italian–Ladin parallel data, which serves as the foundation for subsequent synthetic data generation. To support Italian–Ladin translation, we employ both LLMs and sequence-to-sequence (seq2seq) models. For LLM-based methods, we consider LLaMA 3.1 (8B and 70B) and GPT-4o. The LLaMA models are evaluated under two settings: few-shot learning (FSL) and supervised fine-tuning (FT). In the FSL setting, we use *LLaMA-v3.1-8b-instruct* and *LLaMA-v3.1-70b-instruct* via the DeepInfra[2] API, and GPT-4o via the OpenAI API; an example prompt is provided in Appendix D. For the FT setting, we fine-tune LLaMA models using Together AI[3] with a LoRA adapter (r=32), a batch size of 8, and three training epochs.

For the seq2seq models, we employ MBART-large-50 and NLLB-200-1.3B, which are encoder-decoder architectures designed for multilingual MT. The FT models of both are conducted on an NVIDIA RTX A6000 GPU with a batch size of 8 and over 7 epochs. To accommodate Ladin, an unseen language for both models, we introduce a Ladin-specific language tag in the tokenizer. This modification aids in the identification and processing of Ladin texts during translation tasks, facilitating better handling of the language. Subsequently, we evaluate translation performance using BLEU (computed with SacreBLEU by Post (2018)), ROUGE, and chrF++, providing a comprehensive assessment of translation quality.

3.3 Synthetic data creation

We construct Ladin NLP datasets by translating D_{Ita_SA} and D_{Ita_MCQA} from Italian data to Ladin data. The resulting datasets support multiple tasks, including MT (Italian–Ladin), SA, and MCQA. To ensure translation quality, we use the best-performing MT model, based on its performance on a held-out testing data. Figure 2 outlines the key steps in this data creation process, as following.

Italian-to-Ladin translation. We begin by translating the labeled monolingual Italian datasets of SA and MCQA tasks into Ladin. This translation process is carried out using the best-performing MT model identified in our evaluation, ensuring the high quality of the synthetic Ladin data.

Filtering I. To ensure the semantic quality of the Ladin translation, we apply a filtering step using similarity scores from the Language-Agnostic BERT Sentence Embedding (LaBSE) model (Feng et al., 2022). We selected LaBSE over BERTScore due to its stronger performance in sentence-level semantic similarity tasks. In AD_{Ita_Lad}, LaBSE achieved a higher similarity score (0.68 vs. 0.50), indicating better cross-lingual alignment. This demonstrates LaBSE's effectiveness on low-resource and unseen languages, consistent with the findings reported in the original study. Before filtering, the Italian monolingual dataset had 30,712 SA and 5,000 MCQA instances. This filtering step evaluates the alignment between each Ladin translation and its original Italian counterpart. Specifically, we retain only pairs with a cosine

[2]https://deepinfra.com/
[3]https://www.together.ai/

Training Data	Model	BLEU		Rouge		chrF++	
		Ita→Lad	Lad→Ita	Ita→Lad	Lad→Ita	Ita→Lad	Lad→Ita
AD_{Ita_Lad}	FSL-Llama 3.1 8B	2.87	12.09	18.48	36.79	23.43	36.20
	FSL-Llama 3.1 70B	6.35	21.97	31.81	51.09	31.09	47.26
	FSL-GPT-4o	4.27	22.84	25.86	52.28	27.47	48.23
	FT-Llama 3.1 8B	10.71	15.28	41.56	43.78	35.23	42.04
	FT-Llama 3.1 70B	6.95	22.48	32.97	50.77	31.96	47.96
	Mbart-large-50	10.37	10.91	44.14	41.41	41.73	43.92
	FT-NLLB 1.3 B	17.76	21.41	52.81	49.91	44.60	48.40
	Benchmark (LLM)	3.51	19.44	-	-	25.54	44.69
AD_{Ita_Lad} + SD_{Ita_Lad}	FT-NLLB 1.3 B	18.30	**24.50**	**53.66**	**52.64**	**44.62**	**50.76**
AD_{Ita_Lad} + monolingual data	Benchmark (RBMT)	**18.97**	19.32	-	-	44.13	46.69

Table 1: Comparison of different models based on average translation performance metrics on the test set T. The table is divided into two sections: the upper section reports results using only authentic training data AD_{Ita_Lad}, while the lower section presents results with additional synthetic data.

similarity score $c \geq 0.68$, matching the average semantic similarity observed in the manually created Ladin–Italian parallel sentences in AD_{Ita_Lad}. This threshold helps eliminate semantically inconsistent translations while preserving high-quality pairs for downstream tasks.

Back-translation. To further refine the synthetic dataset, we apply back-translation prior to the second filtering. Ladin translations D_{Lad} are retranslated into Italian, yielding D'_{Ita}, which are then compared to the original Italian data D_{Ita}. This helps identify and remove translations with significant semantic drift, enhancing data quality. Different from previous works (Chauhan et al., 2022; Khatri and Bhattacharyya, 2020) that adopt back-translation approaches, our filtering pipeline is not solely based on BLEU scores. Instead, it incorporates LaBSE semantic similarity in the first filtering stage and METEOR in the second stage to better capture linguistic variation, resulting in a more comprehensive and linguistically informed filtering process.

Filtering II. To finalize the synthetic dataset, we apply the second filtering step using automatic evaluation metrics. Specifically, we compute the BLEU and METEOR scores between the original Italian data D_{Ita} and Italian back-translation D'_{Ita}. We use both metrics because here we compare sentences in the same language, namely Italian and back-translated Italian. BLEU evaluates lexical overlap and fluency, while METEOR captures semantic similarity through recall, synonyms, and morphology. This combination helps us ensure both surface-level and meaning-level quality in the back-translation stage. A translation instance is retained if $BLEU(D_{Ita}, D'_{Ita}) \geq \mu BLEU$ and $METEOR(D_{Ita}, D'_{Ita}) \geq \mu METEOR$, where $\mu BLEU$ and $\mu METEOR$ are the average scores across all pairs. These thresholds help eliminate noisy or inaccurate translations while preserving semantic fidelity. After both filtering stages, the final Italian-Ladin synthetic dataset SD_{Ita_Lad} includes 12,511 SA[4] and 764 MCQA[5] instances.

3.4 Synthetic data evaluation

To assess lexical and syntactic adequacy of the synthetic dataset, we construct a manually translated gold dataset, containing 50 examples[6] from the Italian monolingual dataset to be translated into Ladin by a native speaker of Italian and Ladin. We then compute cosine similarity between Italian and Ladin sentence embeddings in both the synthetic dataset SD_{Ita_Lad} and the gold dataset GD to assess semantic alignment, evaluating whether the machine translation outputs achieve semantic similarity comparable to human translations. Additionally, we adopt the error typology from the Multidimensional Quality Metrics (MQM) framework as qualitative analysis (Lommel et al., 2013), selecting a subset of high-level error dimensions to examine the strengths and limitations of the translations. Specifically, we focus on three dimensions: accuracy, linguistic conventions, and terminology, drawn from the MQM core typology(Lommel et al., 2024), based on the relevance and constraints of our MT process.

[4]SA dataset: https://huggingface.co/datasets/ulinnuha/sentiment_analysis_ladin_italian

[5]MCQA dataset: https://huggingface.co/datasets/ulinnuha/mcqa_ladin_italian

[6]The gold dataset: https://huggingface.co/datasets/ulinnuha/sentiment_analysis_ladin_italian_manual and https://huggingface.co/datasets/ulinnuha/mcqa_ladin_italian_manual

Training Data	Model	BLEU			Rouge			chrF++		
		L	***H***	***Lt***	***L***	***H***	***Lt***	***L***	***H***	***Lt***
AD_{Ita_Lad}	FSL-Llama 3.1 8B	5.12	2.45	1.03	22.87	19.00	13.56	27.36	26.00	16.92
	FSL-Llama 3.1 70B	8.43	7.66	2.96	35.49	33.85	26.09	35.05	35.05	23.17
	FSL-GPT-4o	6.62	4.15	2.03	30.18	26.04	21.36	31.90	30.08	20.42
	FT-Llama 3.1 8B	12.70	9.79	9.64	43.20	40.03	41.44	38.11	36.60	30.98
	FT-Llama 3.1 70B	9.41	7.75	3.70	36.83	34.60	27.47	36.03	35.87	23.99
	Mbart-large-50	11.28	11.34	11.93	43.56	42.75	46.12	39.8	38.66	35.23
	NLLB 1.3 B	18.54	16.85	17.88	**51.59**	51.55	55.29	**46.24**	45.98	41.58
	Benchmark (LLM)	5.54	3.84	1.16	-	-	-	29.03	28.98	18.60
AD_{Ita_Lad} + SD_{Ita_Lad}	FT-NLLB 1.3 B	18.30	17.69	**18.29**	51.40	**52.92**	**56.65**	45.44	46.28	**42.15**
AD_{Ita_Lad} + monolingual data	Benchmark (RBMT)	**20.93**	**19.32**	16.65	-	-	-	**47.65**	**46.58**	38.16

Table 2: Translation performance of various models for Italian → Ladin in 3 test subsets L, H, and Lt. The table is divided into two sections: the upper section reports results using only authentic training data AD_{Ita_Lad}, while the lower section presents results with additional synthetic data.

Training Data	Model	BLEU			Rouge			chrF++		
		L	***H***	***Lt***	***L***	***H***	***Lt***	***L***	***H***	***Lt***
AD_{Ita_Lad}	FSL-Llama 3.1 8B	19.25	12.74	4.27	50.09	38.71	21.56	46.30	38.89	23.41
	FSL-Llama 3.1 70B	30.07	21.55	14.29	62.16	51.60	39.51	56.22	48.66	36.91
	FSL-GPT-4o	29.23	23.38	15.90	61.39	**53.37**	42.07	55.73	50.33	38.64
	FT-Llama 3.1 8B	20.53	17.25	8.05	51.86	44.85	34.62	48.36	44.11	33.66
	FT-Llama 3.1 70B	**30.78**	**23.65**	13.02	**61.91**	52.10	38.29	**57.23**	50.53	36.12
	Mbart-large-50	11.95	12.81	11.84	40.15	42.97	41.12	41.11	42.15	36.81
	NLLB 1.3 B	27.15	19.77	17.31	55.54	48.27	46.03	54.45	49.66	41.09
	Benchmark (LLM)	26.77	21.17	10.37	-	-	-	53.20	48.52	32.36
AD_{Ita_Lad} + SD_{Ita_Lad}	FT-NLLB 1.3 B	30.46	22.71	**20.33**	58.11	50.71	**49.11**	56.95	**51.49**	**43.83**
AD_{Ita_Lad} + monolingual data	Benchmark (RBMT)	21.36	20.27	16.34	-	-	-	50.24	49.08	40.76

Table 3: Translation performance of various models for Ladin → Italian in 3 test subsets L, H, and Lt. The table is divided into two sections: the upper section reports results using only authentic training data AD_{Ita_Lad}, while the lower section presents results with additional synthetic data.

4 Results and Discussion

4.1 Machine translation of Italian and Ladin

Table 1 presents the translation results for both Italian → Ladin and Ladin → Italian across diverse language model configurations. For experiments using AD_{Ita_Lad} as the training dataset, we report results from the LLM benchmark of Frontull and Moser (2024), employing `GPT-3.5-turbo-0125`, for comparison. While Table 1 presents the overall average performance metrics of each translation model evaluated on testing datasets, Tables 2 and 3 offer a more fine-grained analysis. They report detailed metric scores for the individual testing subsets L, H, and Lt, allowing for a closer examination of model behavior across different data segments.

In both translation directions shown in Table 1, the fine-tuned NLLB (FT-NLLB) model achieves significant performance across evaluation metrics. In particular, FT-NLLB performs best on Italian → Ladin translation, reaching a BLEU score close to 18, which indicates relatively good translation accuracy for an extremely low-resource setting. Moreover, FT-NLLB shows consistent performance across all test subsets, as reported in Table 2. Although its performance on Ladin → Italian translation is slightly lower than that of the FSL-GPT-4o in terms of BLEU, FT-NLLB demonstrates robust performance on the Lt (literary domain) subset, where the other models struggle, as shown in Table 3. In general, FT-NLLB achieves the highest chrF++ scores in both translation directions, suggesting stronger overall translation quality.

These findings underscore a persistent challenge in ELRLs. Despite impressive capabilities of LLMs, their performance often suffers in extremely low-resource scenarios due to training biases toward HRLs with abundant data. Thus, LLMs achieve better BLEU scores in Ladin → Italian translation, where the target language is a high-resource language, reflecting the model's stronger proficiency in HRLs. Although Ladin and Italian

belong to the Romance language family, translation into Ladin remains challenging due to its status as the resource-scarce language, including unique vocabulary, dialectal variations, and a lack of direct equivalence to Italian.

The NLLB model demonstrates relatively robust performance, likely benefiting from its multilingual training across 200+ languages, including over 150 LRLs (Team et al., 2022). Notably, the NLLB model was trained on Friulian, a closely related Raeto-Romance language (UNESCO, 2010), which may contribute to its improved generalization on Ladin. Additionally, we evaluate the suitability and effectiveness of the NLLB tokenizer for processing Ladin text. Our analysis reveals an average tokens-per-word ratio of 1.50 for Ladin, compared to 1.39 for Italian. This modest overhead indicates that the NLLB tokenizer segments Ladin reasonably well, suggesting its viability for downstream translation tasks.

4.2 Qualitative error analysis of Ladin translations

To qualitatively evaluate the translation performance of the proposed model (FT-NLLB), we examine several representative examples of translation errors from Italian to Ladin. The analysis follows the MQM-based framework to categorize errors, thereby providing deeper insights into the model's strengths and limitations. To illustrate, we present the following translation examples in Table 4.

Upon closer inspection, the Ladin translation exhibits only minor errors overall. In *Sample 1*, there are two error categories. The phrase "***I á sté*** *döes nöts*" should be corrected to "***I sun stada*** *döes nöts*" to ensure grammatical correctness. The verb of "***Sté***" ("to stay") is intransitive and needs the auxiliary in the past tense. For a first-person singular subject, the correct form is "***I sun stada***." Furthermore, the phrase of "*la vasca* ***da iade***" should be revised to "*la vasca* ***idromassaje***" to more accurately reflect the original Italian expression "*la vasca idromassaggio*." The term "***Iade***" means "journey," indicating an incorrect meaning in the translation. While, the term "***idromassaje***" is the conventional and semantically precise. Then, *Sample 2* indicates that an error does not appear. In *Sample 3*, there is a mistranslation of "*por* ***ester***," which should be translated as "*por* ***podëi gní***." In addition, an omission error also occurs where "*mëssel ester*" should be corrected to "*mëssel* ***pa*** *ester*." Based on the MQM framework, omission and mistranslation errors fall under the accuracy dimension, while grammar-related errors belong to the linguistic convention dimension.

This qualitative analysis further indicates that the model performs robustly on short texts. This can be attributed to the authentic training data AD_{Ita_Lad} used to build FT-NLLB, which consists of very short sentences with an average length of approximately 5 words. In contrast, the Italian monolingual source data used for synthetic generation are substantially longer, with average lengths of more than 45 words for MCQA and more than 70 words for SA. However, these sample analyses in Table 4 illustrate the model's overall ability to generate semantically faithful translations, while also highlighting areas where linguistic precision can be further improved, particularly in longer texts. Overall, the model demonstrates strong potential for translating Ladin effectively.

4.3 Results on Ladin synthetic NLP datasets

Since the FT-NLLB model demonstrates the best translation performance using AD_{Ita_Lad} as training data, we utilize this model to translate Italian into Ladin through the proposed framework.We obtained 30,477 paired entries for the SA dataset after applying the first filtering step based on cosine similarity between Italian and Ladin with the threshold of 0.68. This value is derived from the cosine similarity between both languages in AD_{Ita_Lad}. Subsequently, we perform back-translation from the Ladin SA dataset to Italian. The threshold for the second filtering step is determined based on the average (μ) BLEU and METEOR scores, calculated for each entry by comparing the Italian SA dataset with the back-translated Italian SA dataset. These scores were 33.63 and 0.58, respectively. According to (Cloud, 2025), BLEU scores between 30–40 indicate generally good, understandable translations. The final synthetic SA dataset, consisting of 12,511 entries, statistically is detailed in Table 5. This SA dataset contains an average of 65 and 70 words per entry for Italian and Ladin, respectively. Additionally, the majority of entries in the final dataset, over 9,000 instances, are labeled with positive sentiment.

Concurrently, we obtained 4,651 paired Italian-Ladin entries of the MCQA dataset, following an initial filtering step based on cosine similarity. The thresholds for the second filtering step are also determined using the average BLEU and

No	Italian (Source)	Ladin (Target)	Error
1	*Sono stata a un congresso a Torino e ho passato due notti al Best Western Crystal Palace, un hotel 4 stelle di fianco alla stazione di Porta Nuova. Ho apprezzato molto la vasca idromassaggio e il giornale ogni giorno: La Stampa, con un foglio di sintesi degli argomenti principali.*	*I sun stada a n congres a Torin y i à sté döes nöts tl Best Western Crystal Palace, n hotel da 4 stëres dlungia la staziun de Porta Nuova. I á dër aprijé la vasca da iade y le foliet da vigni de: La Stampa, cun na plata de ressumé di argomënc prinzipai.*	■ Grammar; ■ Mistranslation
2	*Per cosa si utilizza la tecnica epidemiologica della metanalisi?*	*Por ci adoron pa la tecnica epidemiologica dla metanalisa?*	No error
3	*Secondo la Costituzione italiana, per essere eletto Presidente della Repubblica in quale condizione giuridica deve trovarsi il candidato?*	*Aladô dla Costitutziun taliana, por ester lité President dla Republica te ci condiziun iuridica mëssel ester le candidat?*	■ Mistranslation; ■ Omission

Table 4: Qualitative human evaluation of Ladin translations generated by the proposed model.

Feature	Language	
	Italian	Ladin
Average number of words per entry	65	70
Average number of characters per entry	144	348
Positive label count	9,842	9,842
Negative label count	2,669	2,669

Table 5: Summary statistics of the synthetic paired sentiment analysis dataset in Italian and Ladin.

Feature	Language	
	Italian	Ladin
Average number of words per question	18	19
Average number of characters per question	108	97
Average number of words per choices	28	30
Average number of characters per choices	187	174
Frequency of entries with 3 choices	304	304
Frequency of entries with 4 choices	196	196
Frequency of entries with 5 choices	264	264

Table 6: Statistics of the synthetic paired MCQA dataset in Italian and Ladin.

METEOR scores, calculated between the Italian MCQA dataset and the back-translated Italian MCQA dataset, which are 36.58 and 0.62, respectively. The final synthetic MCQA dataset comprises 764 entries, presented in Table 6, with the majority of questions featuring three answer choices. Notably, Italian-Ladin paired datasets of SA and MCQA tasks exhibit a higher average number of words and characters per entry compared to the authentic parallel dataset AD_{Ita_Lad}.

After obtaining the new Italian-Ladin synthetic dataset SD_{Ita_Lad}, we evaluate this new dataset quality by comparing it against the gold dataset GD using cosine similarity to measure semantic alignment between synthetic and manually data. The cosine similarity scores between Italian and Ladin in SD_{Ita_Lad} and GD are 86.61 and 85.75, respectively. This result indicates a high degree of semantic consistency and minimal disparity between synthetic and human-translated datasets.

4.4 Synthetic data assessment for NLP tasks

We combine SD_{Ita_Lad} with AD_{Ita_Lad} to construct new training data for fine-tuning NLLB as the MT model. The previous testing datasets are still used to evaluate translation performance on the new combined datasets, as shown in the lower section of Table 1. Detailed results for the three subsets (L, H, and Lt) are also presented in the lower section of Tables 2 and 3. We compare our results with the augmented synthetic translations generated by the RBMT system proposed by Frontull and Moser (2024), incorporating additional monolingual data in Ladin and Italian. As summarized in the lower section of Table 1, the results demonstrate that incorporating our synthetic data consistently improves translation performance across all evaluation metrics, surpassing the previous benchmark. In both translation directions, the FT-NLLB model outperforms its counterpart. In particular, training with the combined dataset SD_{Ita_Lad} and AD_{Ita_Lad} leads to clear improvements over using AD_{Ita_Lad} alone for Ladin → Italian translation. As shown in Table 3, the model achieves higher BLEU and chrF++ scores across all test subsets, where it previously struggled.

Given that SD_{Ita_Lad} encompasses SA and MCQA tasks in Ladin, we establish a benchmark to evaluate model performance on these tasks. We adopt the FSL approach using LLaMA 3.1 70B as the LLM approach. Appendices E and F provide example prompts used in this approach. In parallel, we evaluate the performance of SA and MCQA tasks using several fine-tuning models, including the Distilbert-base-multilingual-cased (m-DistilBERT) [7] model, the XLM-RoBERTa base model [8], and the mT5-small model[9]. We split each

[7] https://huggingface.co/distilbert/distilbert-base-multilingual-cased

[8] https://huggingface.co/FacebookAI/xlm-roberta-base

[9] https://huggingface.co/google/mt5-small

Model	Ladin				Italian			
	SA		MCQA		SA		MCQA	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
FSL-LLaMA	93.99	96.07	44.20	44.46	97.92	98.18	54.79	54.18
m-DistilBERT	92.28	95.03	36.66	36.69	94.78	96.41	30.73	30.29
XLM-RoBERTa	87.22	93.01	25.21	25.80	97.14	98.08	26.14	26.14
mT5	50.09	69.37	23.50	23.16	49.98	69.25	22.62	21.92

Table 7: Results for SA and MCQA tasks in Ladin, with Italian included for comparison.

SA and MCQA dataset into 80% for training and 20% for testing. For SA, we set `MAX_LEN=256`, `BATCH_SIZE=8`, and train for 3 epochs. For MCQA, we use `MAX_LEN=128`, `BATCH_SIZE=4`, and train for 5 epochs.

Table 7 summarizes the results for SA and MCQA tasks in Ladin[10], with Italian results for comparison. The LLM approach achieves the highest scores on both tasks in terms of balanced accuracy (Acc) and F1-score (F1), although the improvements over other models are marginal. m-DistilBERT also performs reasonably well across both tasks. The results indicate that the Ladin SA dataset is generally well-suited for model training, as most evaluated models perform competitively, with the exception of mT5. In particular, mDistilBERT and LLaMA achieve accuracy above 90.00 in the SA task, demonstrating that the synthetic data is representative and well-aligned with task objectives. These findings suggest that our constructed dataset provides sufficient signal for effective learning in the SA task, and supports generalization across different model architectures. In contrast, performance on the MCQA task remains low across all models, even when evaluated on Italian data. This suggests that the MCQA task is inherently more challenging than the SA task. The difficulty can be attributed to the limited size of the MCQA dataset and the broad topical coverage of the questions, both of which can negatively affect model generalization.

5 Conclusions and Future Work

This work explores various strategies for translating Ladin, an underrepresented language, and demonstrate that our NLLB-based model consistently outperforms counterpart models. Beyond translation, we introduce the first comprehensive Ladin benchmark dataset covering machine translation, sentiment analysis, and question answering, all derived from monolingual Italian resources. Our evaluation shows that while synthetic data enables competitive performance across machine translation and sentiment analysis tasks, further progress is needed to enhance knowledge transfer from high-resource to low-resource languages. This study lays foundational work for future NLP research on Ladin and similarly underrepresented languages.

Future work will investigate knowledge distillation techniques to more effectively transfer knowledge from high-resource languages to low-resource ones, as well as explore advanced methods to better exploit available monolingual Ladin data for improved model performance. We also aim to further support the development of Ladin resources and engage the community in active research on this language by establishing leaderboards based on the created datasets.

Limitations

The Ladin-Italian dataset used in this study primarily consists of short sentences, which may limit the generalizability of the models to longer and more complex sentence structures. Moreover, the limited availability of manually crafted monolingual Italian MCQA datasets restricts the number of usable samples, which may impact the robustness and reliability of the experimental results in the MCQA task.

Acknowledgments

Ulin Nuha acknowledges funding from the Ernst Mach Grant (Grant No. MPC-2024-01571), OeAD-GmbH (Austria), which supported this research at the Data Science Group, University of Innsbruck. We also thank Matthias Winkler, Veronica Rubatscher, and Samuel Frontull for their valuable assistance as Ladin speakers.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via

[10]The implementation can be found at https://github.com/ulinnuhaha/NLP_Ladin.

multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, page 4336–4349.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. Beyond data quantity: Key factors driving performance in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, page 225–239.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 3504–3519.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, page 405–433.

Shweta Chauhan, Shefali Saxena, and Philemon Daniel. 2022. Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low resource languages. *Neural Processing Letters*, 54:1707–1726.

Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. 2024. A novel prompting method for few-shot ner via llms. *Natural Language Processing Journal*, 8(100099).

Google Cloud. 2025. Evaluate models. Accessed September 5, 2025.

Anthony Thomas Connor. 2023. *Ladin Perspectives on Language and Identity in the Central Dolomites of Northern Italy*. Ph.d. thesis, The University of Sheffield, Faculty of Arts and Humanities, School of Languages and Cultures.

Alessandro Desole. 2020. Sentiment analysis on tripadvisor reviews. Accessed: 2025-01-23.

Silvia Erardi, Ronald L. Gardner, and Sara Comploi. 2022. Anglicisms in ladin: Loanwords and local perceptions. *Forum Italicum*, 56(2):272—-306.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 878 – 891.

Samuel Frontull and Georg Moser. 2024. Rule-based, neural and llm back-translation: Comparative insights from a variant of ladin. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, page 128–138.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 2, page 85–91.

Sofía García. 2024. Enhaced apertium system: Translation into low-resource languages of spain spanish–asturian. In *Proceedings of the Ninth Conference on Machine Translation*, page 878–884.

Ali Hatami, Shubhanker Banerjee, Mihael Arcan, Paul Buitelaar, and John Philip McCrae. 2024. English-to-low-resource translation: A multimodal approach for hindi, malayalam, bengali, and hausa. In *Proceedings of the Ninth Conference on Machine Translation*, page 815–822.

Jyotsana Khatri and Pushpak Bhattacharyya. 2020. Filtering back-translated data in unsupervised neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 4334–4339.

Seong Hoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multisource language training in cross-lingual transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, page 712–725.

Wuying Liu, Wei Li, and Lin Wang. 2023. Multiloop incremental bootstrapping for lowresource machine translation. In *Proceedings of Machine Translation Summit XIX*, page 1–11.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas*, volume 2, page 75–94.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

Kaiwen Lu, Yating Yang, Fengyi Yang, Rui Dong, Bo Ma, Aihetamujiang Aihemaiti, Abibilla Atawulla, Lei Wang, and Xi Zhou. 2025. Low-resource language expansion and translation capacity enhancement for llm: A study on the uyghur. In *Proceedings of the 31st International Conference on Computational Linguistics*, page 8360–8373.

Luca Melchior. 2023. Raeto-romance: Romansh, ladin, friulian. *Linguistics*.

Hojjat Mokhtarabadi, Ziba Zamani, Abbas Maazallahi, and Mohammad Hossein Manshaei. 2025. Empowering persian llms for instruction following: A novel dataset and training approach. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, page 31–67.

Ana Alexandra Morim da Silva, Nikit Srivastava, Tatiana Moteu Ngoli, Michael Röder, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2024.

Benchmarking low-resource machine translation systems. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, page 175–185.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of NAACL-HLT 2019*, page 3868–3873.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, page 3501–3516.

Trinh Pham, Khoi M. Le, and Luu Anh Tuan. 2024. Unibridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, page 3168–3184.

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.

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2024. Llms for extremely low-resource finno-ugric languages. *arXiv preprint arXiv:2410.18902*. Cs.CL, version 1.

Pawan Kumar Rajpoot, Nagraj N Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, page 833–838.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Matteo Rinaldi, Jacopo Gili, Maria Francis, Mattia Goffetti, Viviana Patti, and Malvina Nissim. 2024. Multiple choice questions on multiple topics in italian: A calamita challenge. In *CLiC-it 2024: Tenth Italian Conference on Computational Linguistics*.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *arXiv preprint 2411.11295*. Cs.CL, version 1.

Tong Su, Xin Peng, Sarubi Thillainathan, David Guzmán, Surangika Ranathunga, and En-Shiun Annie Lee. 2024. Unlocking parameter-efficient fine-tuning for low-resource language translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 4217–4225.

Felipe Sánchez-Martínez, Juan Antonio Pérez-Ortiz, Aarón Galiano-Jiménez, and Antoni Oliver. 2024. Findings of the wmt 2024 shared task translation into low-resource languages of spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, page 684–698.

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. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, v3. Thu, 25 Aug 2022.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, page 193–202.

Sami Ul Haq, Rudali Huidrom, and Sheila Castilho. 2024. Dcu adapt at wmt24: English to low-resource multi-modal translation task. In *Proceedings of the Ninth Conference on Machine Translation*, page 810–814.

UNESCO. 2010. *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, France.

Gorka Urbizu, Inaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, page 1603–1612.

Giovanni Valer, Nicolò Penzo, and Jacopo Staiano. 2024. Nesciun lengaz lascià endò: Machine translation for fassa ladin. In *Proceedings of CLiC-it 2024: Tenth Italian Conference on Computational Linguistics*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, page 19519–19529.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2024. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13990–14009.

Liang Zhang, Katherine Jijo, Spurthi Setty, Eden Chung, Fatima Javid, Natan Vidra, and Tommy Clifford. 2024. Enhancing large language model performance to answer questions and extract information more accurately. *arXiv preprint 2402.01722*. Cs.CL, version 1.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, page 468–481.

A Additional Experimental Results

This section presents detailed experimental results for machine translation performance using AD_{Ita_Lad} as the sole training data, as reported in the upper section of Table 1. Figure 3 illustrates the chrF++ scores achieved by various models, while Figure 4 presents the corresponding BLEU score comparisons.

B Back-translation Results

As illustrated in Figure 2, we perform a second round of filtering by back-translating Ladin data into Italian and evaluating the results using BLEU and METEOR scores. For the SA dataset, which contains 30,477 samples, the average BLEU and METEOR scores are 33.63 and 0.58, respectively. In the case of the MCQA dataset (4,651 samples), the corresponding averages are 36.58 and 0.62. To ensure transparency, we present representative filtered examples of back-translated Italian and English samples alongside their original counterparts.

Sample 1

Italian: *Sono stata a un congresso a Torino e ho passato due notti al Best Western Crystal Palace, un hotel 4 stelle di fianco alla stazione di Porta Nuova. Ho apprezzato molto la vasca idromassaggio e il giornale ogni giorno: La Stampa, con un foglio di sintesi degli argomenti principali.*

English: *I attended a conference in Turin and spent two nights at the Best Western Crystal Palace, a four-star hotel next to Porta Nuova station. I really appreciated the hot tub and the daily newspaper: La Stampa, which came with a summary sheet of the main topics.*

Back-translated Italian: *Sono stata a un congresso a Torino e ho alloggiato due notti al Best Western Crystal Palace, un hotel a 4 stelle vicino alla stazione di Porta Nova. Ho apprezzato molto la vasca da viaggio e il giornale giornaliero: La Stampa, con un foglio di riassunto degli argomenti principali.*

Sample 2

Italian: *Per cosa si utilizza la tecnica epidemiologica della metanalisi?*

English: *What is the epidemiological technique of meta-analysis used for?*

Back-translated Italian: *A che serve la tecnica epidemiologica della metanalisi?*

Sample 3

Italian: *Secondo la Costituzione italiana, per essere eletto Presidente della Repubblica in quale condizione giuridica deve trovarsi il candidato?*

English: *According to the Italian Constitution, what legal status must a candidate have in order to be elected President of the Republic?*

Back-translated Italian: *Secondo la Fondazione italiana, per essere eletto Presidente della Repubblica in quale condizione giuridica deve essere il candidato?*

Moreover, we present examples of back-translated Italian samples that were discarded after the second filtering stage, as follows.

Sample 4

Italian: *L'albergo non è nemmeno male. Stanze a volte piccole ma senza grosse pretese, colazione decente. L'unica nota stonata è il personale, scontroso, per niente professionale e poco cordiale: sembra che facciano un favore ai clienti. Se non andate di fretta, potete trovare di meglio.*

English: *The hotel isn't even that bad. The rooms are sometimes small, but nothing to complain about if you don't have high expectations, and the breakfast is decent. The only real downside is the staff—gruff, unprofessional, and not very welcoming; they make it feel like they're doing guests a favor. If you're not in a hurry, you can find better options.*

Back-translated Italian: *Campagne piccole ma senza preteste, gustose il pomeriggio. Lunica cosa da sgraffiare è il personale, malleabile, del tutto non professionale e poco cordiale: sembra che*

piacciano ai signori. Se non vado di fretta, puoi trovare qcs. meglio

Sample 5
Italian: *Posizione strategica, zona un po' rumorosa ma va bene lo stesso... stanze belle e servizi all'altezza. La colazione era abbondante e di buona qualità. Da tenere in considerazione per un eventuale nuovo soggiorno a Milano.*
English: *Strategic location, a bit of a noisy area but still acceptable... nice rooms and services up to standard. The breakfast was plentiful and of good quality. Worth considering for a future stay in Milan.*
Back-translated Italian: *Posizione strategica, zona piuttosto cattiva ma interessante... bei cameri e servizio allaperto. La colazione era primaria e di buona qualità. Da considerare per una possibile nuova estasi a Milano.*

Sample 6
Italian: *Hotel di confort notevole. Personale disponibile e attento alle richieste dei clienti. Ottima prima colazione. Anche il ristorante interno è di ottimo livello. Un soggiorno a Napoli da non dimenticare!*
English: *A hotel with outstanding comfort. The staff are helpful and attentive to guests' needs. Excellent breakfast, and the in-house restaurant is also of a very high standard. A stay in Naples not to be forgotten!*
Back-translated Italian: *Albergo con grande comodità. Personale attentivo alle richieste dei clienti. Buon assaggio. Anche il ristorante lì dentro è di buon livello. Una sosta a Napoli indimenticabile!*

C A Few-shot Learning Prompt Template for Correcting Italian Source Data

```
Please review and correct the following Ita
lian texts, ensuring proper Italian grammar,
syntax, and style. Focus on clarity and accu
racy while preserving the original meaning:

    {
[un ottimo hotel in zona centrale,molto par
ticolare e carino,personale gentilissimo,p
rezzi contenuti a due passi dalla stazione
di porta susa,con ristoranti vicinissimi,co
lazione mega abbondante,comprensiva di dol
ce e salato,camere ampie con balcone,unico
neo lo scarico del wc che scrosciava tutta
la notte],
...
[Ottimo hotel situato in posizione comoda
per girare Torino a piedi. Camera e bagno
abbastanza spaziosi e puliti. Grande cort
esia dello staff Colazione abbondante e va
ria. Ottimo rapporto qualitÃ. prezzo Par
cheggio per auto a pagamento molto comodo.
Ci tornerei volentieri! Complimenti]
    }

Return the corrected version in the origi
nal format, as a list (e.g., [xxx], [xxx],
..., [xxx]). Do not include any additional
explanations or the original texts.
```

D A Few-shot Learning Prompt Template in the MT task

```
Here are examples of translations in a JSON
format between Italian and Ladin with the
Val Badia variant:

{
    "translations": [
        {
        "Italian": "è venuta la mia ora",
        "Ladin": "al é gnü mia ora"
        },
        {
        "Italian": "vado dalle cugine!",
        "Ladin": "i vá dales jormanes"
        },
        {
        "Italian": "staccare la luce",
        "Ladin": "destodé la löm"
        },
        ...
        {
        "Italian": "a ottobre inoltrato",
        "Ladin": "d'otober fora"
        }
    ]
   }

Please provide the translation of the fol-
lowing 15 entries in the JSON format, fill-
ing the empty 'Ladin' fields for each entry.
Do not include any additional explanations
or text:

{
    "translations": [
```

```
    {
  "Italian": "imprimere nella mente",
    "Ladin": ""
    },
    {
   "Italian": "temperare la matita",
    "Ladin": ""
    },
    {
  "Italian": "mettere paura a qcn.",
    "Ladin": ""
    },
    ...
    {
  "Italian": "un animale scattante",
    "Ladin": ""
    }
 ]
}
```

E A Few-shot Learning Prompt Template in the SA task

```
Below are Tripadvisor reviews in Ladin
(Val Badia variant) along with their sen
timent labels:

    {
[review: "I sun stá chiló por 7 nes. Le gos
té é bun. Porimpó é i chelins cherdá indorm
edí pormal. ...", label: 1], [review: "Lal
berch é bunorté te na zona chîta dlungia na
plaza. Gní zoruch vigni sëra ê sciöche cia
fé naoasa. ..." label: 0],
...
[review: "I un passé chiló n bel fin dlede
ma hotel nët, personal da orëi bun, bun in
cele gosté y na posiziun ezelënta, impormó
do labela plaza San Marco. ...", label: 0]
    }


Please classify the sentiment for the fol
lowing 10 Tripadvisor reviews in Ladin
(Val Badia variant) as either 0 (Positive)
or 1 (Negative). Fill in the empty 'label'
fields with only 0 or 1. Respond with the
sentiment labels in list format like this:
[x, x, ...]. Do not include any additional
explanations or text.

    {
[review: "Por na vistada y na fuga a Milan
(na mostra, na cörta vijita, na spazirada)
él perfet. ...", label: ], [review: "Le
brunch ne joav nia le prisc. Lhotel é n
pü' dalunc dal Duomo. ...", label: ],
...
[review: "Rezeziun ezelënta, dantadöt le
concierje cun sües racomanaziuns. ...",
label: ]
    }
```

F A Few-shot Learning Prompt Template in the MCQA task

```
Below are multiple-choice questions in Lad
in (Val Badia variant) with 3, 4, or 5 answe
r choices. The correct answer is explicitly
provided as an id number corresponding to
the order of the choices:

    {
[question: aladô dla lege 241/1990, olá é
pa metüda sö la comisciun por l'azes ai do
cumën c aministratifs?, choices: ['pro la
Presidënza dl Consei di Minisć', 'Por vigni
Entité publica dl post', 'pro vigni Entité
publica economica de competënza regionala'
], answer: 0], [question: aladô dla DGR 514
/2009, por PAI y PEI se intendi rispetiva
mënter:, choices: ["Plan d'Assistënza Indi
cisé y Program Etich Individualisé", 'Prog
ram de Asistënza Individuala y Program Edu
catif Individualisé', "Plan d' Assistënza
Individualisé y Plan d'Educaziun Individua
lisé"], answer: 2],
...
[question: aladô de REICAT l'intestaziun
uni forme por na porsona:, choices: ['al
corespogn tresala forma dl inom che vëgn
dant tla pröma ediziundles operes dl au
tur.', 'al se basa söl inom cun chël che
la porsona medema é generalmente identi
fiada.', 'ara ne pó mai ester metüda a
döm da npseudonom.'], answer: 1]
    }

Please answer the questions based on the
available choices, by filling in the empty
'answer' fields with the id number corresp
onding to the order of the choices. Provide
the answers in a list format like this: [x,
x, x, ..., x]. Do not include any additio
nal explanations or text.

    {
[question: na coleziun de figurines é metü
```

```
da adöm da 84 toc, 14 por vigni contignidú.
7 figurines de vigni contignidú é lamincar
ds, i restanc fotocartes. Tan de figurines
de fo to cartes ál pa la coleziun?, choices
: ['42', '43', '44'], answer:],
...
, [question: ci frasa, danter chëses listi
gades, á pa n complemënt de gauja?, choices:
['Laura é piada demez por n iade de plajëi',
'i ápormó cumpré na scincunda por le compli
ann de Marco', "por na taiada de strom s'á
as censur bloché"], answer: ]
   }
```

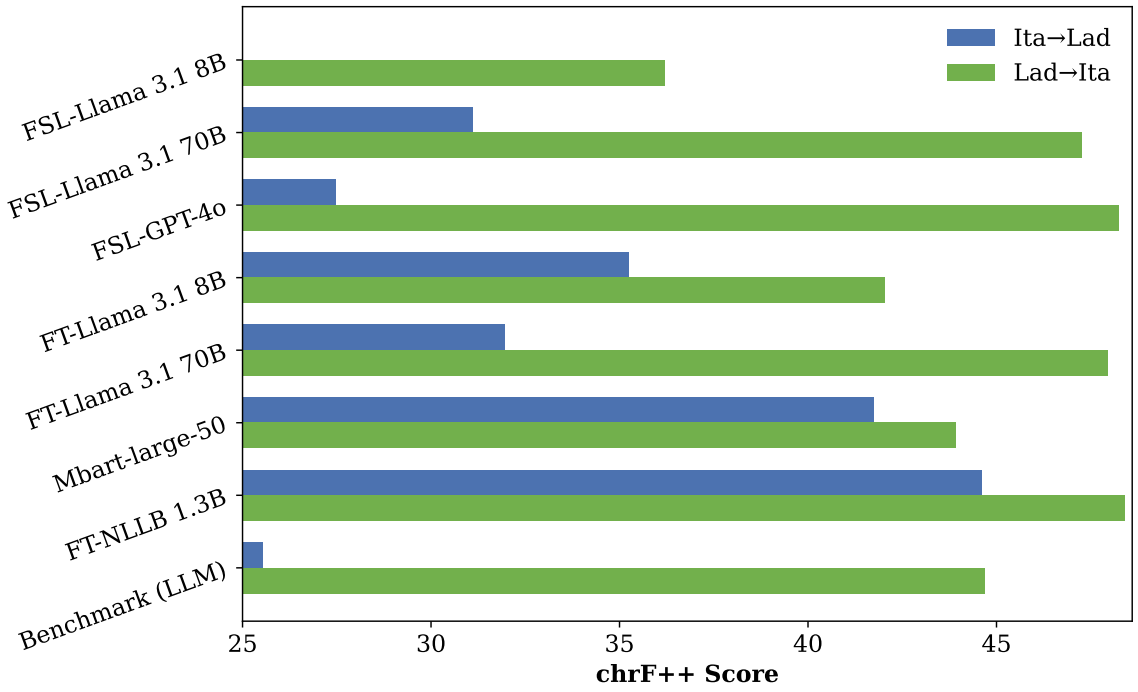

Figure 3: chrF++ scores of the models using AD_{Ita_Lad} as the training data

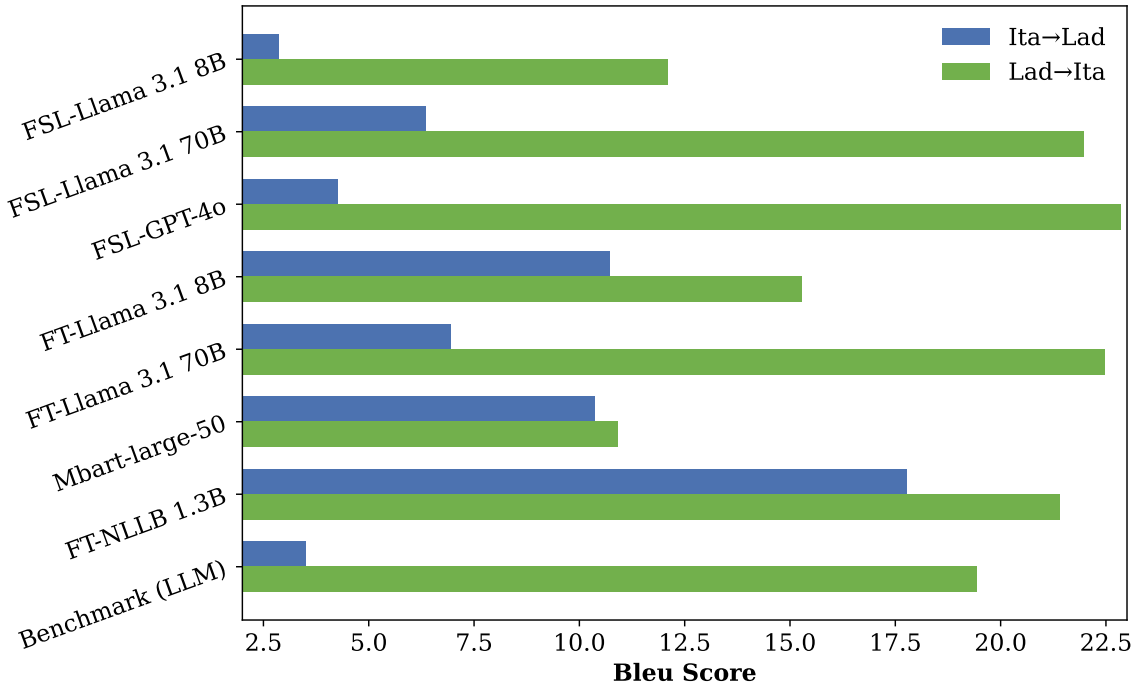

Figure 4: BLEU scores of the models using AD_{Ita_Lad} as the training data