Crowdsourcing Piedmontese to Test LLMs on Non-Standard Orthography

Gianluca Vico and Jindřich Libovický
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
{vico, libovicky}@ufal.mff.cuni.cz

Abstract

We present a crowdsourced dataset for Piedmontese, an endangered Romance language of northwestern Italy. The dataset comprises 145 Italian–Piedmontese parallel sentences derived from Flores+, with translations produced by speakers writing in their natural orthographic style rather than adhering to standardized conventions, along with manual word alignment. We use this resource to benchmark several large language models on tokenization parity, topic classification, and machine translation. Our analysis reveals that Piedmontese incurs a tokenization penalty relative to higher-resource Romance languages, yet LLMs achieve classification performance approaching that of Italian, French, and English. Machine translation results are asymmetric: models translate adequately from Piedmontese into high-resource languages, but generation into Piedmontese remains challenging. The dataset and code are publicly released.

1 Introduction

Piedmontese (ISO 639-3: pms) is a Romance language spoken in the Piedmont region of northwestern Italy. According to Ethnologue (Eberhard et al., 2025), it has fewer than one million speakers and is classified as endangered, with intergenerational transmission in decline.

Existing NLP resources for Piedmontese are limited and predominantly derived from Piedmontese Wikipedia. While useful, these sources largely adhere to standardized orthographic conventions and thus fail to capture the orthographic variations that are common in written Piedmontese. This discrepancy raises the question of how well current language models handle Piedmontese as it is actually written by speakers.

To address this gap, we present a crowdsourced dataset of Italian-to-Piedmontese translations, where annotators were explicitly instructed to write in whichever orthographic style feels natural to them. The source sentences are drawn from the Flores+ dataset (NLLB Team et al., 2024), a multiparallel corpus spanning over 200 languages. We additionally provide manual word alignments between Piedmontese and Italian sentence pairs.

	Flores+ *dev* 114
Ita	Si tratta della maggiore acquisizione nella storia di eBay.
Pms	A l'è la pi granda aquisission ënt la stòria d'ebay
Eng	It is the biggest acquisition in eBay's history.
Fra	C'est la plus grande acquisition de l'histoire d'eBay.

Table 1: Sample from parallel sentences for evaluating machine translation. Annotators translated the Italian sample into Piedmontese. The Italian, French and English samples are originally from Flores+.

Using this data, we evaluate several large language models (LLMs) both intrinsically, through tokenization parity (Petrov et al., 2023) analysis, and extrinsically, on topic classification (using labels from SIB-200; Adelani et al., 2024) and machine translation (MT). Our results indicate that current LLMs exhibit reasonable comprehension of Piedmontese, achieving decent performance in classification and translation from Piedmontese into high-resource languages. However generation into Piedmontese remains substantially more challenging.

We illustrate the data collection procedure in Section 2. In Section 3, we describe the dataset, and in Section 4, we asses LLM performance on Piedmontese. Section 5 presents related datasets and in Section 6 we summarise our findings.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 70–86
March 29, 2026.

The dataset[1] and evaluation code[2] are released under an open-source license.

2 Data Collection

We collected translations via an online questionnaire administered in Italian (see Appendix H), the dominant language in the region, understood by all Piedmontese speakers. Annotators were recruited voluntarily through social media and word of mouth, with no restrictions on repeated participation. To preserve anonymity, we did not track annotator identity across sessions.

The questionnaire comprises three components. First, we elicit demographic and sociolinguistic information, including the annotator's primary language, self-assessed proficiency in Piedmontese, age group, and method of language acquisition. We also ask whether they believe Piedmontese has a standard orthography and, if so, whether it is commonly used. These questions serve both to characterize our annotator population and to contextualize the orthographic variation in the resulting translations.

Second, we present annotators with a randomly selected Italian sentence from Flores+ (NLLB Team et al., 2024) and ask them to translate it into Piedmontese. Crucially, annotators are instructed to write in whatever manner feels natural to them, rather than adhering to any prescribed standard. Translation is optional, so annotators can still complete the other parts of the questionnaire. In this way, we can accommodate annotators who comprehend Piedmontese but do not actively write it. To address the absence of certain diacritics on standard physical keyboards, we provide a substitution scheme (e.g., */:a* for *ä*). Mobile keyboards do not have this issue, and we observed that people either directly use diacritics or use diacritics that can be found on Italian keyboards (*àèéìòù*).

Third, annotators evaluate a translation submitted by a previous participant, viewing both the Italian source and the Piedmontese rendering. This peer review mechanism enables filtering of erroneous or inappropriate submissions and provides an estimate of inter-annotator agreement on translation quality. While the task is subjective, we ask annotators to take into consideration possible variations of the language and of the orthography.

[1] http://hdl.handle.net/11372/LRT-6086

[2] https://github.com/GianlucaVico/CrowdsourcedPiedmontese

3 Dataset Description

We have collected 200 annotations, and 145 of them have valid translations: 68 are from the Flores+ *dev* set, while 77 are from the *devtest* set. 102 samples have been reviewed by at least one annotator, but due to their limited number. We use the reviews only to filter missing or offensive translations.

We organise the collected data in three datasets: 1) the raw list of annotations that can be used for further analysis, 2) a list of parallel sentences for evaluating MT systems, and 3) a list of word-aligned sentences.

3.1 Annotation

Figure 1 shows that most annotators use primarily Italian, and a few use Piedmontese. Other languages include English and Icelandic. The proportion of annotators who submitted a translation is higher among Piedmontese speakers than among Italian speakers. Additionally, most annotators declared themselves to be perfectly or fully proficient in Piedmontese. Most of the annotators are confident in their language knowledge; however, only a small portion considers Piedmontese their native language.

Our questionnaire reached mainly younger people, as shown in Figure 2, but, since this is an endangered language, older people are more likely to speak it.

On average, completing the questionnaire took approximately 7 minutes. People who did not provide a translation took approximately 3 minutes. According to 11% of the annotators, people use standard grammar when writing Piedmontese, while 42% of them disagree. 54% of them think that Piedmontese has a standard grammar, whether it is used or not, and for 25% of the annotators, there is no standard.

3.2 Parallel Sentences

Flores+ contains 2009 samples divided into the *dev* and *devtest* splits. The sentences provided to the annotators are randomly selected, so some of them have multiple translations: three samples from the *devtest* set and one sample from the *dev* set have two translations. The paired sentences have the same overall meaning, but translation quality varies; for example, annotators may use more general terms, summarise a list or remove details. 102 samples have at least one human review, which we

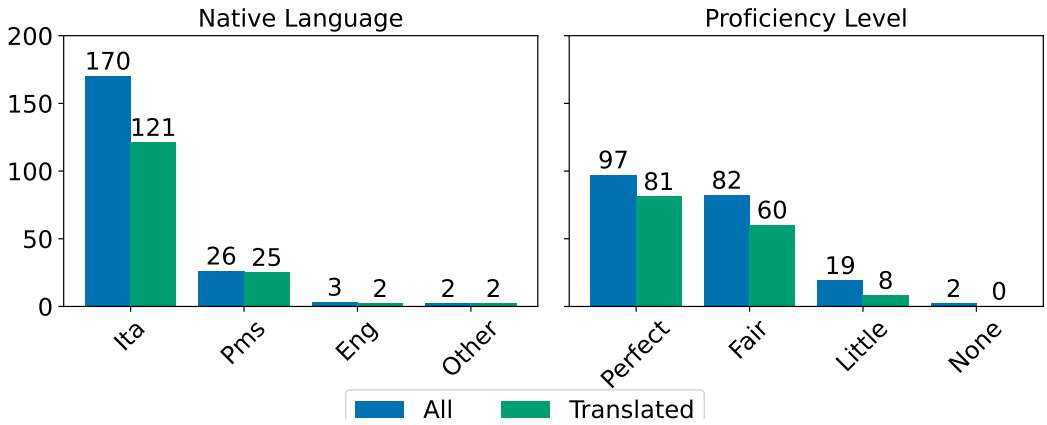

Figure 1: On the left, the main language used by the annotators; Icelandic is included in *Other*. On the right, the self-reported proficiency in Piedmontese. The majority of people uses Italian and self-reports perfect or fair proficiency in Piedmontese.

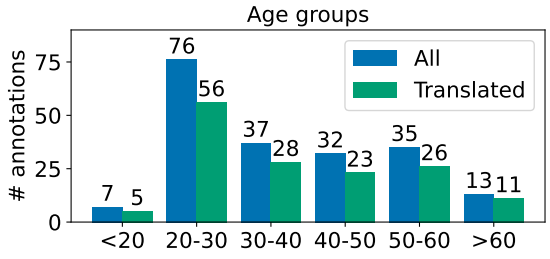

Figure 2: Age distribution of the annotators. Most annotators are 20-30 years old, while older people are more likely to know Piedmontese, but we did not reach them.

used to remove incorrect translations. We present a sample in Table 1: as can be seen, the Piedmontese text may contain incorrect capitalization or missing punctuation. Also, the use of diacritics is inconsistent among annotators.

3.3 Word Aligned Sentences

Due to the limited number of samples, the authors are able to manually word-align the Piedmontese and Italian sentences. We select pairs of corresponding spans in the paired sentences, ensuring that each span is non-overlapping (i.e., each word belongs to at most one span) In this sense, there are cases where, for example, a verb is aligned with a noun because they convey the same meaning, but the sentence structure differs. We use the white space and apostrophe to split words. As an example, *e sull'albero (and on the tree)* consists of three words: *[e][sull'][albero]*. One word can be aligned to multiple words, e.g., *è (is)* is aligned to *a l'è*, and there are unaligned words. However, we do not consider subword alignment. The dataset comprises 3003 spans, with a median of 20 spans per sentence pair. 2902 spans are a single word aligned to another single word. The median number of characters for each span is 5 for both Italian and Piedmontese.

4 Model Evaluation

To assess LLM performance on Piedmontese, we first evaluate tokenizer parity (Petrov et al., 2023): this provides an estimate of the costs in tokens of processing Piedmontese compared to other languages. Then, we use the aligned dataset to investigate whether models can find corresponding words between Piedmontese and Italian. Finally, we use topic classification and machine translation as downstream tasks for evaluation. In topic classification, models need to be able to understand Piedmontese, while in machine translation, they also have to generate Piedmontese. The downstream tasks are evaluated in a zero-shot setup.

We consider the following open-weight models from HuggingFace: Llama 3.3 70B[3] (Grattafiori et al., 2024), Gemma 3 27B[4] (Gemma Team, 2025), Qwen 3 30B[5] (Qwen Team, 2025), EuroLLM 9B[6] (Martins et al., 2025), Tower Plus 9B[7] (Rei et al., 2025); and the closed-source models: Gemini[8] and GPT[9]. Besides Piedmontese and Italian, we also

[3] meta-llama/Llama-3.3-70B-Instruct
[4] google/gemma-3-27b-it
[5] Qwen/Qwen3-30B-A3B-Instruct-2507
[6] utter-project/EuroLLM-9B-Instruct
[7] Unbabel/Tower-Plus-9B
[8] gemini-2.5-flash-preview-09-2025
[9] gpt-4o-mini

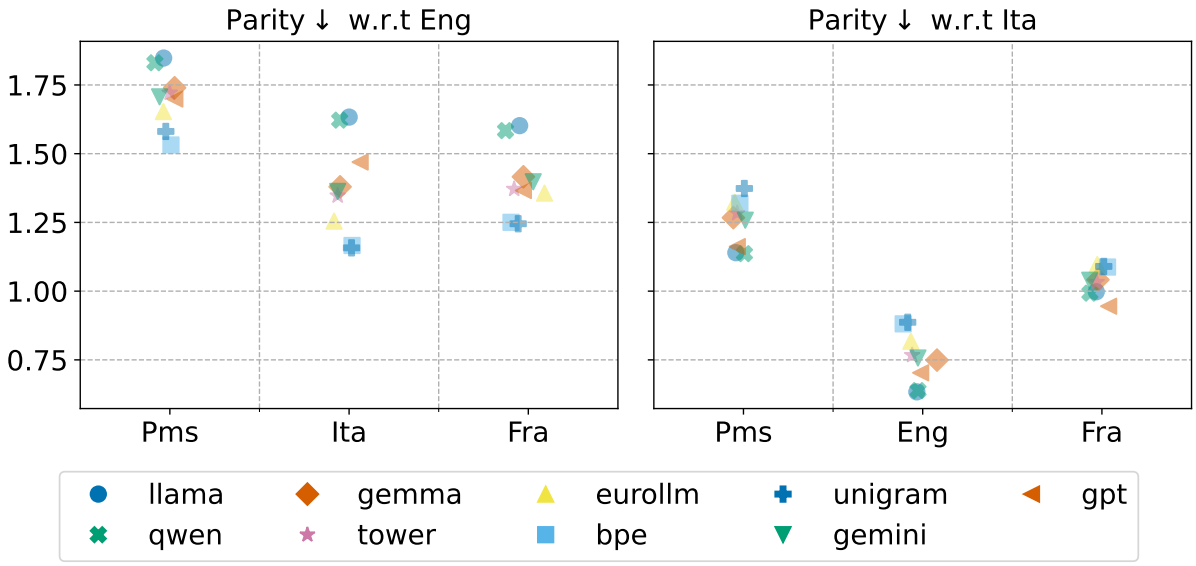

Figure 3: Parity scores with respect to English and Italian. Piedmontese has worse parity compared to the other languages; however, it is closer to one when compared to Italian.

include French, as it is the high-resource language closest to Piedmontese, other than Italian, and English, due to its widespread availability. Because the data is limited and we do not perform any parameter search, we evaluate the model on the combined *dev* and *devtest* sets. The hyper-parameters for the experiments are listed in Appendix F.

4.1 Tokenizer Parity

As shown by Ahia et al. (2023), low-resource languages are often overtokenized, resulting in higher costs and worse performance compared to high-resource languages. We evaluate the tokenizer parity (Petrov et al., 2023) for the LLMs, UnigramLM, and BPE from SentencePiece to estimate the number of extra tokens required to process the same sentence in Piedmontese.

We train the SentencePiece tokenizers using English, Italian, French, and Piedmontese data from the Glot500 Corpus (Imani et al., 2023), with 100k samples for each language and a vocabulary size of 32,000.

We average the parity of each sample, computed as:

$$p_t(s_{\text{tgt}}, s_{\text{ref}}) = \frac{|t(s_{\text{tgt}})|}{|t(s_{\text{ref}})|}$$

where t is a tokenization function that produces a list of tokens, s_{tgt} is a sentence in the target language, and s_{ref} is the corresponding sentence in the reference language. As reference languages, we use English and Italian. A parity close to one indicates that the tokenizer produces a similar number of tokens for the source and target languages, whereas values greater than one indicate that the target language is over-tokenized.

In Figure 3, we report the parity scores of the models. Piedmontese has worse (i.e., higher) parity than the other languages, which means that using LLMs with Piedmontese is more computationally expensive. Training the tokenizer on Piedmontese can help, as BPE and UnigramLM have lower parity compared to English. However, overall, the models yield comparable results, and closed models do not have an advantage over the open-weight models. In Appendix D, we report the exact parity scores for the different setups.

Model	$F_1 \uparrow$	Precision↑	Recall↑
Eflomal	.774	.817	.735
SimAlign	.589	.726	.496

Table 2: Alignment scores of eflomal and SimAlign.

4.2 Word Alignment

We use eflomal (Östling and Tiedemann, 2016) trained on our dataset as a baseline and compare with the unsupervised SimAlign (Jalili Sabet et al., 2020) with XLM-RoBERTa (Conneau et al., 2020) with subwords. XLM-RoBERTa is a multilingual model, but Piedmontese was not explicitly included in the training data. We use the same evaluation script from SimAlign, which reports precision, re-

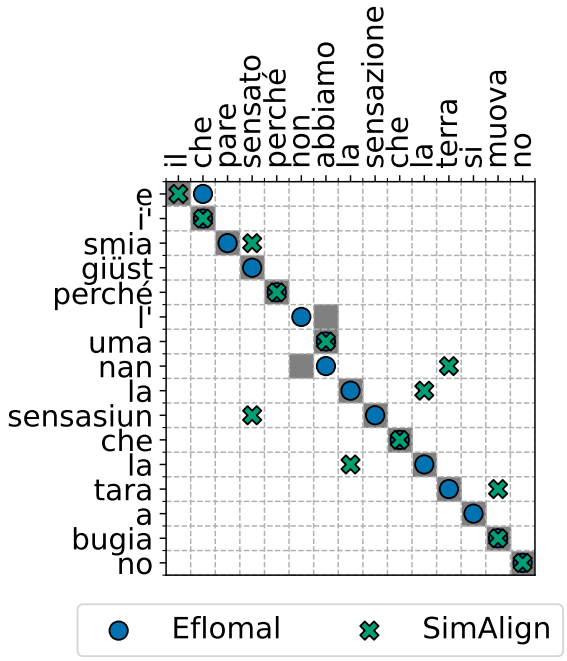

Figure 4: This is a sample alignment. The gray background is the reference alignment, while eflomal alignment is represented by the blue circles and SimAlign one by the green crosses. The English translation of the sentence is *This seems sensible, because the Earth doesn't feel as if it's moving, does it?*

call, F_1, and alignment error rate (AER) defined as:

$$\text{prec} = \frac{|A \cap P|}{|A|}, \ \text{rec} = \frac{|A \cap S|}{|S|},$$
$$F_1 = \frac{2\text{prec} \cdot \text{rec}}{\text{prec} + \text{rec}}, \ \text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

where A are the system alignment, S the sure reference alignment and P the possible reference alignments. However, our annotations do not include possible alignment and so AER is simply $1 - F_1$ (see Appendix E).

The results in Table 2 show that eflomal achieves better scores than SimAlign, which relies on the language model representations. The scores of SimAlign are comparable to those that its authors observed for English-Hindi alignment, indicating that the model produces reasonable alignments despite not being trained on Piedmontese. This indicates that while the XLM-RoBERTa representations are sufficient for generating zero-shot alignment, statistical methods still yield better results.

Additionally, the effect of (sub)words that are identical between Italian and Piedmontese and are therefore easier to align is unknown. We show an example of alignment in Figure 4, where the reference alignment is mostly monotonic. SimAlign seems to align words in the wrong position (e.g., *la* aligned to the wrong *la*), while eflomal might align words that occur together (e.g., the negation *nan* with the verb *abbiamo*).

4.3 Topic Classification

We use SIB-200 (Adelani et al., 2024) to evaluate the models on topic classification with our data. SIB-200 uses sentences from Flores+, so it is possible to obtain labels for the Piedmontese sentences. The dataset contains 7 classes: *science/technology*, *travel*, *politics*, *sports*, *health*, *entertainment*, and *geography*, but some sentences are labelled as *uncategorized* and are excluded from our experiments. In total, 37 sentences from the *dev* set and 38 from the *devtest* set have a label. We use the same set of sentences for all four languages.

In Figure 5, we report the F_1 scores of the models on the different languages. We note that, while scores for Piedmontese are generally lower than for the other languages, they are still comparable, meaning that models are able to understand the language. Smaller models, such as EuroLLM, exhibit worse performance: in particular, EuroLLM struggles to follow instructions and, in French, often generates all labels or overly long explanations. Tower has a larger drop in performance in Piedmontese, but it is still able to solve the task despite its focus on machine translation. Closed models behave similarly to the larger open-weight models. See Appendix B for the exact values and additional metrics and Appendix A for the prompts used.

4.4 Machine Translation

We test zero-shot machine translation, including both *Pms*→*X* and *X*→*Pms*. From Figure 6, models have similar chrF++ (↑) scores when translating from the different languages to Piedmontese. Moreover, all languages achieve similar chrF++ scores when translating to Italian, including Piedmontese. However, translating from Piedmontese to French or English is worse than in other languages. While we cannot directly compare the target languages, *X*→*Pms* has noticeably lower scores.

Given that *X*→*ita* has comparable results for all languages, we use Italian as a pivot by first translating from the source language to Italian, and then to the target language. From Table 3, the pivot strategy improves translations up to +2.15 chrF++ in the *Pms*→*X* direction and +1.22 chrF++ in the

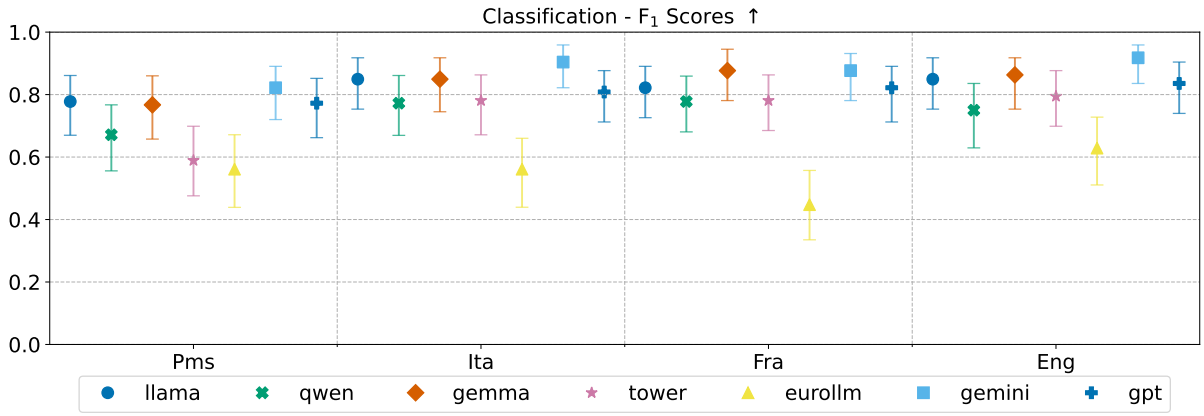

Figure 5: Comparison on the F1 scores of the models in the topic classification task. We perform bootstrapping to compute the confidence interval of the scores.

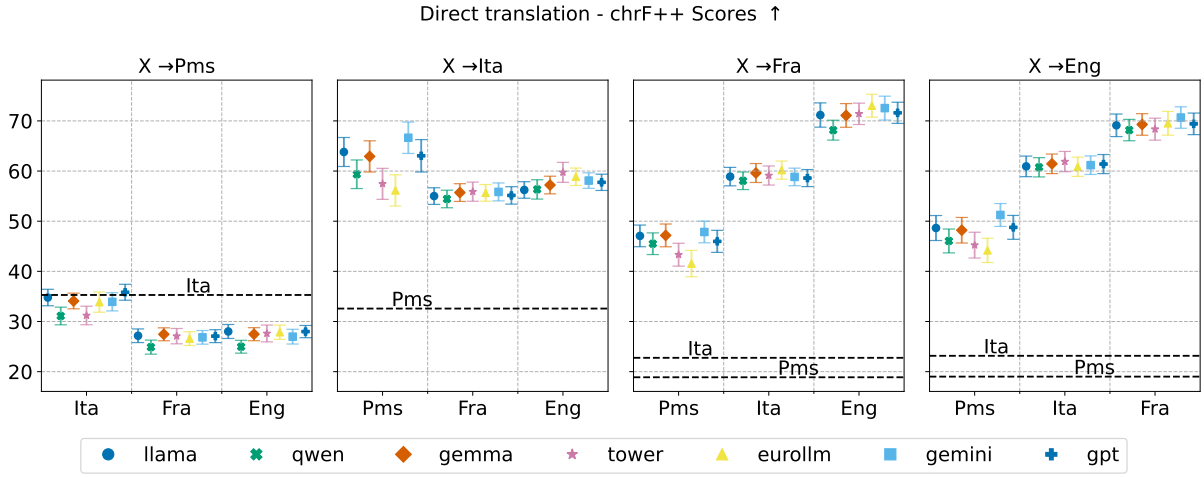

Figure 6: chrF++ scores of the different models. Each subplot shows the target language, while the sources languages are on the x-axis. The dotted horizontal lines indicate the scores obtained when using the reference text in a given language as if it were the translation.

X→Pms direction.

However, evaluating *X→Pms* is particularly challenging, because models might produce Piedmontese that is correct but different from the reference, which does not use standard orthography, and surface-level metrics such as chrF++ penalize this. Machine-learned metrics like COMET (Rei et al., 2020) can improve this, but they need training data, which is not available. In Appendix C, we report additional metrics, including COMET (without fine-tuning on Piedmontese). Moreover, we observe that the Italian sentences are closer to the Piedmontese references than what the models are generating, as shown by the horizontal lines in Figure 6. This can also explain why LLMs are able to understand Piedmontese. In Table 4, we show some translation examples between Italian and Piedmontese from different models. See Appendix C for the exact values and additional metrics and Appendix A for the prompts used.

5 Related Work

The data for this work is derived from Flores+ (NLLB Team et al., 2024), which is an evaluation benchmark for machine translation. It contains 2009 sentences, each translated into more than 200 languages. It does not include Piedmontese, but it includes geographically close Italian regional languages, such as Ligurian and Lombard. There are other datasets that contain Piedmontese data, such as Wikipedia (68k samples) and Wikisource (4k samples) (Wikimedia Foundation), and Glot500 (Imani et al., 2023) (226k samples), which derive from the Piedmontese portion of Wikipedia, and Tatoeba (Tiedemann, 2020) (800 samples), which contains sentences annotated by volunteers. How-

	Dev	Target			
		Pms	Ita	Fra	Eng
Source	Pms	-	58.80	42.48	45.95
	Ita	32.62	-	57.83	60.98
	Fra	26.45	54.68	-	70.89
	Eng	27.02	57.60	72.25	-
	Pms_{Pivot}	-	-	44.62	46.91
	Fra_{Pivot}	27.67	-	-	67.71
	Eng_{Pivot}	28.07	-	67.96	-

	Devtest	Target			
		Pms	Ita	Fra	Eng
Source	Pms	-	62.15	47.42	47.77
	Ita	33.23	-	60.26	61.37
	Fra	26.60	56.11	-	67.90
	Eng	26.97	57.92	70.61	-
	Pms_{Pivot}	-	-	49.57	49.91
	Fra_{Pivot}	27.11	-	-	65.11
	Eng_{Pivot}	27.85	-	67.63	-

Table 3: Average chrF++ scores of the models on the two sets and all directions. Note that only columns are comparable. *Pivot* refers to the experiments that use Italian as a pivot for the translation.

Flores+ *dev* 114	
It is the biggest acquisition in eBay's history.	
Pms	A l'è la pi granda aquisission ënt la stòria d'ebay
Ita	Si tratta della maggiore acquisizione nella storia di eBay.
Pms→Ita	
EuroLLM	È la più grande acquisizione nella storia di eBay
Gemini	È la più grande acquisizione nella storia di eBay
Ita→Pms	
Gemma	A l'é la pì gròssa acquisission an la stòria d'eBay.
GPT	A l'é la piò gròssa aquisission an sla storia ëd eBay.

Table 4: Translation examples. The *Ita→Pms* translations are understandable, but have different spellings than the reference and across models. The *Pms→Ita* translations are correct, although, phrased differently than the reference translation.

ever, these datasets contain a more standardised version of Piedmontese, which differs from what people might use in real life.

Datasets derived from CommonCrawl[10], like C4 (Raffel et al., 2020), FineWeb2 (Penedo et al., 2025), CulturaX (Nguyen et al., 2024), and Oscar (Ortiz Su'arez et al., 2020), also contain some Piedmontese, but correctly identifying a low-resource language is challenging, and false positives can affect the results.

Another project with the objective of collecting language data, including Piedmontese, is AlpiLinK (Rabanus et al., 2023–). AlpiLinK collects crowdsourced spoken data of various regional languages in the Alpine regions of Italy and contains 5442 Piedmontese sentences.

Ramponi and Casula (2023) propose DIATOPIT, a corpus of social media posts written in different local languages of Italy or using regional Italian. The corpus includes 288 Piedmontese samples and, similarly to our work, does not assume a standard orthography but focuses on the languages as actually written by people.

6 Conclusion

In this paper, we presented a crowdsourced dataset for Piedmontese, whose main characteristic is the non-standard orthography. The dataset can be used for further research on the annotators' demographics, machine translation, and word alignment. Furthermore, we highlight how Piedmontese is at a disadvantage in many popular NLP models, showing that it has higher parity compared to related languages. We then test several LLMs to investigate their performance on topic classification and machine translation tasks. We note that models are able to understand Piedmontese, although they perform worse than in other languages; the scores are still comparable. However, generation still remains challenging.

7 Limitations

This work presents several limitations. Firstly, the selection of annotators is biased because it relies on social media, and people who speak the language may not be accessible. This also influences how the translations are written, because some characters are easier to type on a smartphone keyboard than

[10] https://commoncrawl.org/

on a physical one or with pen and paper. Additionally, we do not track which variant of Piedmontese the annotators use, but we consider Piedmontese to be what the annotators themselves refer to as Piedmontese. Then, the annotators are mostly Italian native speakers, since Italian is the national language, and the number of samples is extremely small. We focus on Piedmontese in Italy and do not consider, for example, Piedmontese spoken in Argentina. The task involves translating from Italian, which can result in translationese. Also, in the questionnaire, we use the terms *orthography* and *grammar* interchangeably to make it easier to understand.

Acknowledgments

This research was supported by the Czech Science Foundation project 25-16242S. The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ. We thank the annotators who contributed to this work. GV thanks friends and relatives and the Instagram pages piemontays, Spurgatocn and Abitare il Piemontese for sharing the questionnaire to a larger public.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28th edition. SIL International, Dallas, TX.

Gemma Team. 2025. Gemma 3.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Pedro Javier Ortiz Su'arez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Aleksandar Petrov, Emanuele La Malfa, Philip H.S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Stefan Rabanus, Anne Kruijt, Birgit Alber, Ermenegildo Bidese, Livio Gaeta, Gianmario Raimondi, Paolo Benedetto Mas, Sabrina Bertollo, Serena Bissolo, Angelica Bonelli, Dario Capelli, Jan Casalicchio, Raffaele Cioffi, Patrizia Cordin, Michele Cosentino, Silvia Dal Negro, Ilaria Driussi, Alexander Glück, Joachim Kokkelmans, and 10 others. 2023–. AlpiLinK. German-Romance language contact in the Italian Alps: documentation, explanation, participation. `https://alpilink.it`. Ongoing project.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Alan Ramponi and Camilla Casula. 2023. Diatopit: A corpus of social media posts for the study of diatopic language variation in italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, page 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

A Prompts

Topic classification. The system prompt is `"You are a helpful assistant that classifies the following sentence into one of the following categories: science/technology, travel, politics, sports, health, entertainment, geography. Do not add any explanations."`

The user prompt is `"Is this a piece of news regarding "science, technology, travel, politics, sports, health, entertainment, or geography"? TEXT."`, where `TEXT` is the sentence we are classifying. For Tower we did not use the system prompt.

Machine translation. The system prompt is `You are a helpful assistant that translates the following sentence from SRG to TGT. Do not add any explanations.`

The user prompt is `Translate the following SRC source text to TGT:\nSRC: SENTENCE\nTGT: "`. `SRG` and `TGT` are the name of the source and target language. `SENTENCE` is the sentence to translate. For the pivot experiments, the first step translates to Italian, while the second translates from Italian.

B Classification Results

Tables 5 to 8 show the scores on the text classification task.

	Metric: F_1			
Model	Pms	Ita	Fra	Eng
Llama	0.778 (0.047)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.671 (0.054)	0.772 (0.047)	0.778 (0.046)	0.750 (0.051)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.579 (0.058)	0.524 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.772 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 5: F_1 of the different models on the classification task. In parenthesis the STD of the score.

Metric: Precision				
Model	Pms	Ita	Fra	Eng
Llama	0.789 (0.047)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.686 (0.054)	0.778 (0.047)	0.789 (0.046)	0.761 (0.051)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.583 (0.058)	0.528 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.778 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 6: Precision of the different models on the classification task. In parenthesis the STD of the score.

Metric: Recall				
Model	Pms	Ita	Fra	Eng
Llama	0.767 (0.048)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.658 (0.054)	0.767 (0.047)	0.767 (0.047)	0.740 (0.052)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.575 (0.058)	0.521 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.767 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 7: Recall of the different models on the classification task. In parenthesis the STD of the score.

Metric: Accuracy				
Model	Pms	Ita	Fra	Eng
Llama	0.767 (0.048)	0.849 (0.041)	0.822 (0.044)	0.849 (0.040)
Qwen	0.658 (0.054)	0.767 (0.047)	0.767 (0.047)	0.740 (0.052)
Gemma	0.767 (0.050)	0.849 (0.042)	0.877 (0.037)	0.863 (0.041)
Tower	0.589 (0.057)	0.781 (0.049)	0.781 (0.049)	0.795 (0.046)
Eurollm	0.575 (0.058)	0.521 (0.058)	0.479 (0.057)	0.671 (0.054)
Gemini	0.822 (0.044)	0.904 (0.034)	0.877 (0.037)	0.918 (0.031)
Gpt	0.767 (0.048)	0.808 (0.045)	0.822 (0.045)	0.836 (0.042)

Table 8: Accuracy of the different models on the classification task. In parenthesis the STD of the score.

C Machine Translation Results

Tables 9 to 20 show the scores for the direct MT task (standard deviation in parenthesis), while Tables 21 to 26 show the scores with pivoting.

Pms → Ita				
Model	BLEU	chrF++	TER	COMET
Llama	42.910 (3.720)	63.800 (2.890)	44.080 (3.650)	0.933 (0.003)
Qwen	37.490 (3.730)	59.360 (2.850)	48.890 (3.810)	0.932 (0.003)
Gemma	42.090 (4.260)	62.920 (3.110)	45.960 (4.300)	0.935 (0.003)
Tower	34.450 (4.100)	57.450 (3.090)	56.690 (5.850)	0.931 (0.003)
Eurollm	32.290 (3.690)	56.140 (3.110)	56.130 (4.300)	0.909 (0.009)
Gemini	46.290 (4.310)	66.640 (3.130)	42.450 (5.390)	0.937 (0.002)
Gpt	42.290 (4.170)	63.040 (3.230)	44.300 (4.060)	0.932 (0.004)

Table 9: Translation results from Pms to Ita.

Pms → Fra				
Model	BLEU	chrF++	TER	COMET
Llama	21.700 (2.400)	47.070 (2.160)	70.330 (3.470)	0.908 (0.005)
Qwen	20.420 (2.500)	45.510 (2.160)	71.000 (2.920)	0.905 (0.005)
Gemma	21.550 (2.640)	47.170 (2.270)	70.680 (3.490)	0.908 (0.005)
Tower	17.770 (2.310)	43.320 (2.280)	77.660 (3.690)	0.906 (0.005)
Eurollm	16.670 (2.500)	41.550 (2.630)	79.670 (3.580)	0.901 (0.006)
Gemini	21.550 (2.690)	47.850 (2.160)	71.460 (4.840)	0.908 (0.005)
Gpt	20.670 (2.410)	45.980 (2.200)	69.420 (3.110)	0.906 (0.005)

Table 10: Translation results from Pms to Fra.

Pms → Eng				
Model	BLEU	chrF++	TER	COMET
Llama	20.270 (2.760)	48.640 (2.490)	71.780 (4.460)	0.901 (0.004)
Qwen	19.180 (2.690)	46.060 (2.380)	74.230 (3.790)	0.900 (0.004)
Gemma	19.970 (2.810)	48.200 (2.550)	73.510 (4.220)	0.900 (0.004)
Tower	17.540 (2.620)	45.230 (2.570)	80.070 (4.900)	0.900 (0.004)
Eurollm	16.080 (2.750)	44.180 (2.420)	82.440 (4.930)	0.894 (0.006)
Gemini	22.800 (2.680)	51.240 (2.280)	68.290 (3.800)	0.899 (0.004)
Gpt	20.940 (2.720)	48.770 (2.380)	71.950 (3.760)	0.899 (0.005)

Table 11: Translation results from Pms to Eng.

Model	Ita → Pms BLEU	chrF++	TER	COMET
Llama	6.980 (1.290)	34.780 (1.640)	83.220 (2.550)	0.842 (0.008)
Qwen	5.250 (1.540)	31.100 (1.760)	93.290 (7.820)	0.843 (0.007)
Gemma	6.270 (1.310)	34.090 (1.560)	84.580 (2.560)	0.841 (0.008)
Tower	4.560 (1.240)	31.200 (1.850)	99.970 (12.440)	0.843 (0.007)
Eurollm	6.250 (1.830)	33.870 (2.020)	88.590 (12.490)	0.844 (0.007)
Gemini	7.280 (1.470)	33.930 (1.820)	85.080 (3.030)	0.833 (0.009)
Gpt	6.800 (1.370)	35.830 (1.590)	82.320 (2.530)	0.848 (0.006)

Table 12: Translation results from Ita to Pms.

Model	Ita → Fra BLEU	chrF++	TER	COMET
Llama	33.500 (2.550)	58.890 (1.840)	54.750 (2.930)	0.943 (0.003)
Qwen	32.380 (2.510)	58.060 (1.750)	55.300 (2.940)	0.943 (0.003)
Gemma	34.260 (2.720)	59.610 (1.890)	53.520 (3.150)	0.942 (0.003)
Tower	34.100 (2.660)	59.110 (1.900)	53.900 (2.900)	0.943 (0.003)
Eurollm	34.700 (2.450)	60.190 (1.810)	53.000 (2.990)	0.944 (0.003)
Gemini	33.100 (2.710)	58.830 (1.750)	55.880 (3.090)	0.943 (0.003)
Gpt	32.400 (2.320)	58.600 (1.700)	55.270 (2.710)	0.942 (0.003)

Table 13: Translation results from Ita to Fra.

Model	Ita → Eng BLEU	chrF++	TER	COMET
Llama	32.520 (2.910)	60.930 (2.050)	52.970 (3.430)	0.941 (0.003)
Qwen	32.050 (2.790)	60.750 (1.920)	52.360 (3.110)	0.942 (0.003)
Gemma	33.410 (2.860)	61.440 (1.950)	51.580 (3.100)	0.941 (0.003)
Tower	33.900 (2.800)	61.890 (2.020)	51.610 (3.160)	0.942 (0.003)
Eurollm	32.970 (2.810)	60.850 (1.920)	52.460 (3.140)	0.941 (0.003)
Gemini	32.400 (2.670)	61.170 (1.850)	53.240 (3.350)	0.941 (0.003)
Gpt	32.730 (2.620)	61.390 (1.900)	52.430 (3.090)	0.941 (0.003)

Table 14: Translation results from Ita to Eng.

Model	Fra → Pms BLEU	chrF++	TER	COMET
Llama	3.000 (1.120)	27.160 (1.360)	95.790 (2.510)	0.835 (0.008)
Qwen	2.410 (1.110)	24.890 (1.400)	101.020 (6.610)	0.838 (0.007)
Gemma	2.960 (0.860)	27.450 (1.290)	94.370 (2.300)	0.831 (0.009)
Tower	2.930 (1.120)	27.080 (1.530)	97.820 (7.040)	0.836 (0.008)
Eurollm	2.810 (1.100)	26.590 (1.370)	95.590 (4.060)	0.836 (0.008)
Gemini	3.350 (0.990)	26.840 (1.360)	96.200 (2.670)	0.829 (0.009)
Gpt	3.280 (1.000)	27.080 (1.290)	95.270 (2.170)	0.835 (0.008)

Table 15: Translation results from Fra to Pms.

Model	Fra → Ita BLEU	chrF++	TER	COMET
Llama	27.890 (2.270)	55.010 (1.660)	58.320 (2.660)	0.952 (0.003)
Qwen	26.810 (2.450)	54.430 (1.750)	59.620 (2.710)	0.952 (0.003)
Gemma	29.270 (2.530)	55.710 (1.770)	57.120 (2.650)	0.952 (0.003)
Tower	29.680 (2.690)	55.890 (1.910)	57.400 (2.800)	0.952 (0.003)
Eurollm	29.200 (2.500)	55.650 (1.660)	57.270 (2.670)	0.952 (0.003)
Gemini	29.520 (2.650)	55.840 (1.790)	57.980 (3.020)	0.952 (0.003)
Gpt	27.680 (2.390)	55.140 (1.740)	57.640 (2.700)	0.951 (0.003)

Table 16: Translation results from Fra to Ita.

Model	Fra → Eng BLEU	chrF++	TER	COMET
Llama	46.510 (3.530)	69.120 (2.250)	37.150 (3.270)	0.947 (0.002)
Qwen	45.320 (3.510)	68.170 (2.130)	38.340 (3.400)	0.947 (0.002)
Gemma	46.410 (3.390)	69.290 (2.150)	36.160 (3.170)	0.947 (0.002)
Tower	45.030 (3.330)	68.350 (2.190)	37.830 (3.170)	0.945 (0.002)
Eurollm	47.380 (3.810)	69.520 (2.370)	36.880 (3.480)	0.945 (0.002)
Gemini	48.790 (3.640)	70.670 (2.140)	34.530 (3.290)	0.947 (0.002)
Gpt	46.460 (3.550)	69.410 (2.150)	36.770 (3.100)	0.947 (0.002)

Table 17: Translation results from Fra to Eng.

Eng → Pms				
Model	BLEU	chrF++	TER	COMET
Llama	3.170 (1.140)	28.010 (1.390)	94.430 (2.370)	0.824 (0.009)
Qwen	2.340 (1.200)	24.970 (1.280)	98.110 (3.870)	0.828 (0.008)
Gemma	2.460 (0.840)	27.470 (1.300)	94.370 (2.370)	0.825 (0.008)
Tower	3.170 (1.140)	27.620 (1.680)	99.740 (9.360)	0.830 (0.008)
Eurollm	3.050 (1.070)	27.860 (1.420)	94.830 (6.580)	0.828 (0.008)
Gemini	3.130 (0.890)	26.980 (1.470)	95.730 (4.030)	0.818 (0.010)
Gpt	2.290 (0.920)	27.990 (1.220)	92.130 (1.940)	0.830 (0.007)

Table 18: Translation results from Eng to Pms.

Eng → Ita				
Model	BLEU	chrF++	TER	COMET
Llama	29.850 (2.110)	56.230 (1.650)	54.900 (2.390)	0.953 (0.003)
Qwen	29.920 (2.560)	56.340 (1.930)	56.200 (2.790)	0.952 (0.003)
Gemma	30.770 (2.510)	57.210 (1.760)	54.380 (2.430)	0.953 (0.003)
Tower	34.510 (3.070)	59.740 (1.990)	51.570 (2.610)	0.953 (0.003)
Eurollm	33.590 (2.530)	58.860 (1.740)	52.000 (2.500)	0.953 (0.003)
Gemini	31.660 (2.320)	58.100 (1.550)	53.300 (2.430)	0.953 (0.003)
Gpt	31.240 (2.340)	57.750 (1.620)	52.740 (2.320)	0.952 (0.003)

Table 19: Translation results from Eng to Ita.

Eng → Fra				
Model	BLEU	chrF++	TER	COMET
Llama	52.800 (3.640)	71.170 (2.420)	33.280 (2.970)	0.949 (0.002)
Qwen	48.010 (3.130)	68.150 (1.980)	38.380 (2.820)	0.948 (0.002)
Gemma	52.420 (3.510)	71.090 (2.360)	33.750 (2.920)	0.947 (0.003)
Tower	53.020 (3.360)	71.410 (2.130)	34.360 (2.800)	0.949 (0.002)
Eurollm	55.550 (3.680)	73.030 (2.280)	32.350 (3.040)	0.950 (0.002)
Gemini	55.030 (3.690)	72.550 (2.380)	32.150 (3.060)	0.945 (0.003)
Gpt	53.010 (3.260)	71.610 (2.110)	33.280 (2.890)	0.950 (0.002)

Table 20: Translation results from Eng to Fra.

Pms → Fra, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	23.580 (2.450)	49.090 (2.220)	68.000 (3.290)	0.927 (0.002)
Qwen	19.890 (2.460)	45.140 (2.180)	72.630 (3.190)	0.927 (0.002)
Gemma	23.740 (2.820)	49.450 (2.290)	68.140 (3.570)	0.927 (0.002)
Tower	20.380 (2.460)	46.070 (2.150)	73.850 (3.860)	0.927 (0.002)
Eurollm	20.210 (2.850)	44.900 (2.540)	74.080 (3.840)	0.927 (0.002)
Gemini	25.600 (2.410)	50.780 (2.150)	64.740 (2.930)	0.927 (0.002)
Gpt	23.930 (2.270)	49.320 (2.090)	66.070 (2.940)	0.927 (0.002)

Table 21: Translation results with pivoting from Pms to Fra.

Pms → Eng, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	21.780 (2.910)	49.550 (2.450)	68.830 (3.870)	0.917 (0.002)
Qwen	19.340 (2.730)	47.160 (2.400)	70.800 (3.610)	0.917 (0.002)
Gemma	23.630 (3.050)	50.770 (2.500)	66.010 (4.070)	0.917 (0.002)
Tower	18.650 (2.740)	47.080 (2.430)	75.450 (4.590)	0.917 (0.002)
Eurollm	18.140 (2.940)	45.500 (2.570)	77.720 (4.310)	0.917 (0.002)
Gemini	25.320 (2.730)	53.280 (2.360)	64.010 (3.740)	0.917 (0.002)
Gpt	23.490 (2.950)	50.760 (2.440)	66.860 (3.950)	0.917 (0.002)

Table 22: Translation results with pivoting from Pms to Eng.

Fra → Pms, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	3.990 (1.180)	28.780 (1.430)	92.600 (2.440)	0.836 (0.008)
Qwen	2.720 (1.120)	26.220 (1.430)	99.710 (7.940)	0.838 (0.007)
Gemma	3.760 (1.060)	28.640 (1.430)	92.540 (2.410)	0.834 (0.008)
Tower	2.460 (1.040)	26.440 (1.590)	107.140 (15.430)	0.837 (0.008)
Eurollm	2.870 (1.060)	26.960 (1.480)	95.120 (5.920)	0.835 (0.008)
Gemini	3.880 (1.110)	28.000 (1.550)	94.160 (2.800)	0.829 (0.009)
Gpt	3.580 (1.080)	29.110 (1.400)	93.090 (2.280)	0.838 (0.007)

Table 23: Translation results with pivoting from Fra to Pms.

Fra → Eng, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	42.730 (3.390)	66.800 (2.190)	39.970 (3.580)	0.952 (0.001)
Qwen	42.940 (3.490)	65.990 (2.240)	40.340 (3.410)	0.952 (0.001)
Gemma	42.950 (3.260)	66.630 (2.020)	39.080 (3.130)	0.952 (0.001)
Tower	40.260 (3.210)	64.990 (2.100)	42.240 (3.450)	0.952 (0.001)
Eurollm	41.720 (3.450)	65.570 (2.250)	42.000 (3.750)	0.952 (0.001)
Gemini	43.330 (3.650)	67.260 (2.150)	39.590 (3.400)	0.952 (0.001)
Gpt	42.820 (3.490)	66.820 (2.320)	39.970 (3.490)	0.952 (0.001)

Table 24: Translation results with pivoting from Fra to Eng.

Eng → Pms, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	3.770 (1.050)	28.910 (1.410)	92.680 (2.490)	0.826 (0.008)
Qwen	2.200 (0.980)	26.200 (1.370)	96.920 (3.500)	0.828 (0.008)
Gemma	3.560 (1.040)	29.210 (1.420)	91.640 (2.340)	0.826 (0.008)
Tower	2.850 (1.110)	27.250 (1.710)	104.880 (11.890)	0.828 (0.008)
Eurollm	3.250 (1.180)	28.190 (1.410)	91.060 (1.830)	0.827 (0.008)
Gemini	3.970 (1.170)	28.850 (1.570)	92.650 (2.720)	0.822 (0.009)
Gpt	3.650 (1.100)	29.600 (1.380)	90.910 (2.120)	0.828 (0.008)

Table 25: Translation results with pivoting from Eng to Pms.

Eng → Fra, Pivot				
Model	BLEU	chrF++	TER	COMET
Llama	47.630 (3.130)	67.830 (2.170)	37.390 (2.810)	0.956 (0.001)
Qwen	42.230 (2.970)	64.330 (1.970)	42.110 (2.750)	0.956 (0.001)
Gemma	47.460 (3.390)	67.640 (2.240)	38.090 (2.950)	0.956 (0.001)
Tower	47.240 (3.190)	67.840 (2.090)	38.470 (2.730)	0.956 (0.001)
Eurollm	48.710 (3.400)	68.500 (2.310)	37.040 (3.090)	0.956 (0.001)
Gemini	50.750 (3.590)	69.800 (2.440)	35.350 (3.080)	0.956 (0.001)
Gpt	48.130 (3.350)	68.090 (2.330)	37.010 (3.030)	0.956 (0.001)

Table 26: Translation results with pivoting from Eng to Fra.

D Parity Results

Tables 27 to 30 show the tokenizer parity scores with respect to the different languages. Note that scores with respect to different languages are not comparable.

Parity w.r.t Pms			
Model	Ita	Fra	Eng
Llama	0.905	0.898	0.569
Qwen	0.909	0.897	0.576
Gemma	0.816	0.850	0.609
Tower	0.806	0.830	0.616
Eurollm	0.784	0.858	0.639
Bpe	0.783	0.847	0.687
Unigram	0.752	0.817	0.664
Gemini	0.820	0.853	0.618
Gpt	0.883	0.833	0.618

Table 27: Parity scores with respect to Piedmontese.

Parity w.r.t Ita			
Model	Pms	Fra	Eng
Llama	1.140	0.998	0.634
Qwen	1.136	0.993	0.638
Gemma	1.267	1.041	0.749
Tower	1.284	1.033	0.767
Eurollm	1.324	1.097	0.818
Bpe	1.320	1.088	0.881
Unigram	1.373	1.090	0.887
Gemini	1.258	1.040	0.757
Gpt	1.162	0.945	0.703

Table 28: Parity scores with respect to Italian.

Parity w.r.t Fra			
Model	Pms	Ita	Eng
Llama	1.164	1.028	0.640
Qwen	1.167	1.034	0.649
Gemma	1.243	0.983	0.726
Tower	1.269	0.993	0.751
Eurollm	1.232	0.935	0.753
Bpe	1.235	0.941	0.816
Unigram	1.286	0.941	0.823
Gemini	1.235	0.984	0.734
Gpt	1.258	1.086	0.751

Table 29: Parity scores with respect to French.

Model	Parity w.r.t Eng Pms	Ita	Fra
Llama	1.848	1.633	1.602
Qwen	1.831	1.622	1.584
Gemma	1.740	1.380	1.416
Tower	1.721	1.347	1.371
Eurollm	1.653	1.255	1.357
Bpe	1.532	1.166	1.250
Unigram	1.581	1.158	1.245
Gemini	1.707	1.363	1.398
Gpt	1.699	1.470	1.366

Table 30: Parity scores with respect to English.

E Alignment Metrics

We do not use possible reference alignments, so $|P| = |S|$. Assuming that $|A \cap S|$, $|A|$, and $|S|$ are not empty, F_1 can be rewritten as:

$$\mathrm{F}_1 = \frac{2\mathrm{prec} \cdot \mathrm{rec}}{\mathrm{prec} + \mathrm{rec}} = \frac{2\dfrac{|A \cap P|}{|A|} \cdot \dfrac{|A \cap S|}{|S|}}{\dfrac{|A \cap P|}{|A|} + \dfrac{|A \cap S|}{|S|}} =$$
$$= \frac{2\dfrac{|A \cap S|}{|A|} \cdot \dfrac{|A \cap S|}{|S|}}{\dfrac{|A \cap S|}{|A|} + \dfrac{|A \cap S|}{|S|}} =$$
$$= \frac{2|A \cap S|}{|A||S|} \cdot \frac{|A||S|}{|A| + |S|} = \frac{2|A \cap S|}{|S| + |A|}$$

And AER as:

$$\mathrm{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} = 1 - \frac{2|A \cap S|}{|S| + |A|}$$

Therefore $\mathrm{AER} = 1 - \mathrm{F}_1$

F Hyper-parameters

For the translation task and the classification, we use greedy decoding and we generate at most 100 tokens, which is sufficient for the ground-truth labels. The closed source models do not use reasoning. The open-weight model are used with the Transformers 4.57.1 text generation pipeline, while the closed models are used through OpenRouter.

We run eflomal (version 2.0.0) with its default parameters, then we symmetrize the alignment with fast align atools with *grow-diag-final-and*. SimAlign is run with the following arguments:

- Model: xlm-roberta-base
- Tokenizer type: bpe
- Distortion: 0
- Layer: 8
- Matching method: itermax

G Computational Resources and Costs

The total cost for Gemini was $0.62 and $0.17 for GPT. The provider has a zero data retention policy. The other experiments were run on a local cluster with up to 2 NVIDIA H100 with 95GB VRAM and 60GB RAM.

H Questionnaire

The questionnaire is in Italian. Here we show the original version and in brackets the English translation. The questionnaire introduction explains to the user the goal of the project and emphasizes that it is not about evaluating the user and that it is anonymous.

INFORMAZIONI SU DI TE E SUL PIEMONTESE
[About you and Piedmontese]

Quale lingua utilizzi di più quotidianamente (scuola, lavoro, in giro, ecc.)?
[Which language do you use most on a daily basis?]
Seleziona una lingua dall'elenco. Se scegli "Altro", specifica la lingua nel campo di testo.
[Select a language from the list. If you choose "Other," specify the language in the text field.]
{Seleziona una lingua: Italiano, Francese, Spagnolo,Tedesco, Inglese, Rumeno, Arabo, Macedone, Albanese, Piemontese, Preferisco non rispondere, Altro}
{Select a language: Italian, French, Spanish, German, English, Romanian, Arabic, Macedonian, Albanian, Piedmontese, Prefer not to reply, Other}

Quanto bene parli il piemontese?
[How well do you speak Peidmontese]
Seleziona una delle opzioni che meglio descrive la tua conoscenza del piemontese.
[Select one of the options that best describes your knowledge of Piedmontese.]

- Niente o quasi, solo qualche parola *[Nothing or almost nothing, just a few words]*
- Poco, conosco alcune espressioni, ma faccio fatica a esprimere frasi nuove *[Not much, I know some expressions, but I struggle to express new sentences]*
- Abbastanza, ma a volte lo mischio con l'italiano (o la lingua che uso principalmente) *[Quite a bit, but sometimes I mix it with Italian (or whatever language I use mostly)]*
- Perfettamente o quasi, riesco a esprimere praticamente tutto *[Perfectly or almost perfectly, I can express practically everything]*

Secondo te, il piemontese ha una grammatica e ortografia ben definita ("questa parola si scrive così", "questo verbo si coniuga così")?
[In your opinion, does Piedmontese have a well-defined grammar and spelling ("this word is written like this", "this verb is conjugated like this")?]
Seleziona una delle opzioni che meglio descrive la tua opinione.
[Select one of the options that best describes your opinion.]
{D'accordo, Neutrale, In disaccordo}

Quando le persone scrivono in piemontese usano questa grammatica?
[When people write in Piedmontese, do they use this grammar?]
Seleziona una delle opzioni che meglio descrive la tua opinione.
[Select one of the options that best describes your opinion.]
{D'accordo, Neutrale, In disaccordo}

Da chi hai imparato il piemontese?
[Where did you learn Piedmontese from?]
Puoi selezionare più opzioni. Se selezioni "Altro", puoi specificare.
[You can select multiple options. If you select "Other," you can specify.]

- Nonni *[Grand parents]*
- Genitori *[Parents]*
- Parenti *[Relatives]*
- Amici o colleghi *[Friends or colleagues]*
- Altro *[Other]*

Qual è la tua fascia d'età?
[What is your age range?]
Fai 30 anni tra 4 giorni? Seleziona "Tra 20 e 30" — Hai compiuto 40 l'altro ieri? Seleziona "Tra 40 e 50"
[Are you turning 30 in 4 days? Select "Between 20 and 30." — Did you turn 40 the day before yesterday? Select "Between 40 and 50."]

- Meno di 20 *[Less than 20]*
- Tra 20 e 30 *[Between 20 and 30]*

- Tra 30 e 40 *[Between 30 and 40]*
- Tra 40 e 50 *[Between 40 and 50]*
- Tra 50 e 60 *[Between 50 and 60]*
- Più di 60 *[More than 60]*
- Preferisco non rispondere *[I prefer not to answer]*

TRADUZIONE
[Translation]

Come scriveresti questa frase in piemontese?
[How would you write this sentence in Piedmontese?]
Linee guida:

- Se non sai come tradurla non scrivere nulla. Se vuoi puoi riprovare e il questionario dovrebbe proporre una frase casuale diversa. *[If you don't know how to translate it, don't write anything. If you want, you can try again and the questionnaire should suggest a different random sentence.]*
- Non usare traduttori automatici (Google Translate, ecc.). *[Do not use automatic translators (Google Translate, etc.).]*
- Non aggiungere spiegazioni (no "la traduzione è:", "... (vuole anche dire ...)") o diverse traduzioni possibili (no: "... (che vuole dire ...)", "... opzione 1/opzione 2 ..."). *[Do not add explanations (no "the translation is:", "... (also means ...)") or multiple possible translations (no: "... (which means ...)", "... option 1/option 2 ...").]*
- Può essere che alcune parole siano difficilmente traducibili. Scrivile come le scriveresti tu. *[Some words may be difficult to translate. Write them as you would.]*
- Puoi chiedere aiuto ai nonni. *[You can ask your grandparents for help.]*
- Accenti e simboliche magari non sono sulla tastiera ('a' come esempio). Da telefono puoi tenere premuta una lettera per vedere le opzioni disponibili. *[Accents and symbols may not be on the keyboard ('a' as an example). On your phone, you can press and hold a letter to see the available options.]*
  à: /'a, á: /"a, â: /^a, ã: ~a, ä: /:a, ą: /,a, å: /.a, å: /°a, ă: /=a, ø: //o

In italiano *[In Italian]*
`Sample`
In piemontese *[In Piedmonetese]*
`Text field`

VALUTAZIONE
[Evaluation]

Come valuteresti la seguente traduzione?
[How would you rate the following translation?] Considera possibili variazioni del piemontese (ad esempio di qualcuno di Torino o di Verduno). La traduzione è stata fatta da un'altro utente e presentata senza alcuna modifica.
[Consider possible variations in Piedmontese (for example, someone from Turin or Verduno). The translation was done by another user and presented unchanged.]

In italiano *[In Italian]*
`Sample`
In piemontese *[In Piedmonetese]*
`Sample`

- Interamente corretta o quasi *[Completely correct or almost]*

- Probabilmente corretta, l'avrei scritta in altro modo *[Probably correct, I would have written it differently]*
- Parzialmente corretta *[Partially correct]*
- Totalmente sbagliata o quasi *[Totally wrong or almost]*
- Non lo so *[I do not know]*
- Risposta mancante, offensiva o non pertinente *[Missing, offensive or irrelevant response]*