

GRIAL-TA participation in the WMT26 General Translation Shared Task: SalamandraTA7bFFT

Marta Cozzini^{1*} Antoni Oliver¹ Gonzalo López-Sánchez¹
Mercè Vázquez¹ Patricia Morales-Hurtado¹ Sergi Alvarez-Vidal²

¹Universitat Oberta de Catalunya (UOC), Barcelona, Spain
{mcozzini, aoliverg, glopezsanch, mvazquezga, pmoraleshu}@uoc.edu

²Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
sergi.alvarez@uab.cat

Abstract

This work describes the Grial-TA team submission to the constrained track of WMT2026 General Machine Translation Task (Kocmi et al., 2026) for the following language pairs: English to-Icelandic, English-to-Ladin and English-to-Ligurian. Our submission leverages full fine-tuning of the SalamandraTA-7b-instruct model. Two separated models were fine-tuned: one for Icelandic and one for Ligurian and Ladin. To address the problem of data scarcity inherent to such low-resource languages, we implemented a data augmentation strategy based on two complementary approaches: generating synthetic back-translation into English for all language pairs, and incorporating parallel data from related Romance languages for Ladin and Ligurian.

1 Introduction

Machine translation (MT) is a subfield of computational linguistics that focuses on the conversion of text or speech from one language to another using computational algorithms. Since the launch of large language models (LLMs) to the general public, MT has evolved exponentially and improved translation to the point where it resembles professional human translation (Sizov et al., 2024). However, MT into low-resource languages still faces some challenges due to a lack of parallel corpora and the predominance of high-resource languages used in training (Shu et al., 2024). In order to advance research in this area of knowledge, the Grial-TA team has participated in the constrained track of the WMT2026 General Machine Translation shared task. The task aims to improve and evaluate current MT methods. Among the goals of this edition, our contribution focuses on investigating the translation of low-resource, morphologically rich languages. Specifically, we selected English-to-Ladin and English-to-Ligurian as low-resource lan-

guage pairs, along with English-to-Icelandic which is both low-resource and morphologically rich. To train our system on these language pairs, we created parallel corpora from Wikipedia, aligning with the overall goal of building publicly available resources for low-resource machine translation.

2 Resources

In this section, we specify the training resources, processing tools and data augmentation strategies used in our workflow. Specifically, Subsection 2.1 introduces the pipeline for extracting corpora from Wikipedia, while Subsection 2.2 covers the creation of a language identification tool specifically designed for Romance languages. Finally, Subsection 2.3 details the data augmentation strategy that we implemented to address data scarcity.

2.1 Corpus creation from Wikipedia

Wikipedia has long been recognized as a fundamental resource for building large-scale, multilingual corpora due to its vast size, broad domain coverage, and permissive licensing. It is also a particularly valuable source for building corpora for low-resource languages, which is precisely the case for Ligurian and Ladin, languages for which data availability is extremely limited.

We used the tool `corpus-from-wikipedia` (Segués et al., 2026) to create parallel corpora aligning English with both our primary target languages (Ladin, Ligurian and Icelandic) and related Romance languages (Friulian, Lombard, Piedmontese, Romansh and Latin). These auxiliary resources serve to address severe data scarcity for Ladin and Ligurian by complementing the training set with parallel corpora from related languages.

In the following subsections we describe the key steps we followed to build the corpora (see Figure 1 for a step-by-step illustration).

* Corresponding author.

Text extraction

The first step consisted in processing the entire Wikipedia dump (released 01/04/2026) and extracting all articles for both languages of each pair. To prevent data imbalance between English and less-resourced languages, we restricted the extraction only to articles shared by both languages, retaining only the English pages with a corresponding counterpart in the target language.

Sentence segmentation and deduplication

The extracted page content was segmented into sentences using a set of discrete rules defined in a Segmentation Rules eXchange (SRX) file. Since our SRX file does not support Icelandic or Ligurian, we used Stanza (Qi et al., 2020) to segment the articles. Given that Ladin is also not supported in our SRX file or other external Python libraries we segmented the text using the generic rules of our SRX file. After segmentation, the sentences were deduplicated to guarantee a final corpus of unique segments.

Sentence alignment

The subsequent step in the pipeline was sentence alignment for each pair of languages. It should be noted that equivalent Wikipedia pages in two different languages are independent texts and are not necessarily translations of each other, although some content may overlap. Consequently, creating a parallel corpus from this source is not a classical text alignment task. Instead, it is a process of bitext mining, which involves searching for parallel segments within comparable, non-parallel documents. First, the deduplicated sentences were filtered by character length, restricting them to segments between 10 and 200 characters. We used the pre-trained Language-agnostic BERT Sentence Embedding (LaBSE) model, implemented via the sentence-transformers library (Reimers and Gurevych, 2020) to encode the segments into a vector space. We then used Faiss (Johnson et al., 2021) to conduct a bidirectional exact k -NN search, retrieving the $k = 4$ closest matches for each segment. We calibrated the candidate scores using a ratio-based margin function, meaning a pair received a high score only if the two segments exhibited strong mutual similarity relative to their local neighborhoods. Finally, the highest-scoring candidates from the bidirectional search were aggregated and globally ranked. The final parallel sentences were extracted if their margin score exceeded a

threshold of 1.0. If the source or target segment had already been assigned to a higher-scoring pair, the candidate was discarded. This ensured that only the most reliable alignments were retained for the final corpus.

Language verification and rescoring

Since the alignment may contain errors, such as segments not in the expected language, we applied the rescoring algorithm described by Oliver and Álvarez (2023) to clean the aligned corpus from cross-lingual noise. This algorithm uses `lid.176.bin` (Joulin et al., 2016), a language identification model based on `fastText`¹. This model is trained on large multilingual datasets and supports 176 languages, however, Ligurian and Ladin are not included, a challenge detailed in section 2.2.

Filtering and selection of final segments

Following the rescoring step, we selected the rescored segments based on two key metrics: language detection confidence and similarity between segments. The defined threshold for these two metrics is 0.75, that is, segments need to have a language detection score and a similarity score of 0.75 or higher. Segments that met both thresholds were retained to form the final corpus. The final Wikipedia corpora statistics are presented in Table 1; the "Selected" column represents the final corpus size for each pair of languages.

2.2 Language detection in Ligurian and Ladin

As mentioned in section 2.1, the `fastText` language identification model `lid.176.bin` does not support Ligurian or Ladin. Given that reliable language identification is essential for building parallel corpora from Wikipedia, we created a new language detector tailored to those two languages, alongside the additional Romance languages introduced above. Some of these languages are already supported in the `fastText` model, which, however, is not always accurate in recognizing them. Furthermore, considering the languages that are most likely to appear in the Wikipedia texts we intend to process, we also included Italian, French, German, Spanish, and English. In total, the final detector covers 12 languages: Ladin, Ligurian, Piedmontese, Lombard, Romansh, Friulian, Latin, Italian, French, English, German, and Spanish.

When building a language detector it is important to keep the training data balanced,

¹<https://fasttext.cc/>

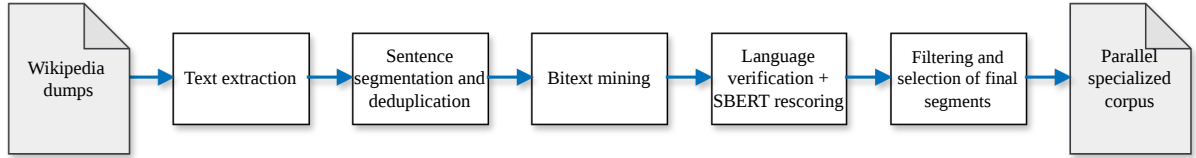


Figure 1: Corpus from Wikipedia construction pipeline.

Language Pair	Unique (SL)	Unique (TL)	Unaligned (SL)	Unaligned (TL)	Aligned	Selected
Icelandic-English	947,697	4,323,323	417,499	3,792,373	530,983	50,134
Ligurian-English	80,491	901,359	40,713	861,567	39,792	616
Ladin-English	277,879	2,398,258	209,946	2,329,220	69,056	1,315
Lombard-English	306,334	2,055,737	170,560	1,919,976	135,780	5,196
Piedmontese-English	324,868	1,791,042	211,093	1,677,193	113,856	4,539
Friulian-English	53,188	349,484	28,015	324,302	25,182	899
Romansh-English	117,890	208,027	65,859	155,852	52,176	1,268
Latin-English	1,775,569	3,033,232	1,167,809	2,425,153	608,159	17,667

Table 1: Corpora statistics.

i.e., each language should be represented by a similar number of examples to prevent the model from becoming biased towards languages with more representation. In our case, we aimed for around 100,000 segments per language. For major languages such as English, Spanish, French, Italian, or German, we sourced the data from Wikipedia, following the workflow described in section 2.1. We extracted 300,000 random sentences from each monolingual corpus, labeled them by language using the fastText language identification model and randomly sampled 100,000 in the corresponding language from the resulting ones. We followed the same process with Piedmontese, given that it is already a language supported in said model. Although Romansh and Lombard are also supported, the model fails to detect these languages accurately in most cases. For this reason, rather than extracting segments specifically labeled as Romansh/Lombard, we opted to filter out the most prevalent majority languages. After manually reviewing a substantial portion of the data, we established a negative filtering threshold: any segment classified as English, German, or French with a confidence score of 0.60 or higher was removed from the data. Friulan was not supported by the original detector, but followed the same process as we did for Lombard and Romansh. For Ligurian and Ladin we did not source the data from Wikipedia. Instead, we decided to use data that were made available by the organizers of the shared task. For Ladin we used the Sfrontull-la_usc_valbadia_loresmt24-1-1ld.1ld dataset. It contains

237,343 segments from which we randomly selected 100,000. For Ligurian we used the Conseggioligure-linc-1-lij_Latn.lij_Latn dataset. After segmentation, it resulted in a final corpus of 84,611 segments. For Latin we created a corpus from Perseus², a free digital library of classical texts, from which we also randomly selected 100,000 segments.

Subsequently, the compiled data was partitioned into training and test sets following a 90:10 split. When evaluated on this held-out data, our language identification model achieved a precision at 1 (P@1) and recall at 1 (R@1) of 0.967, demonstrating robust overall performance. Both metrics are computed at k=1, considering only the top-ranked prediction. A class-level evaluation on the test set reveals that major languages achieve high scores (around 0.99 F1-score), indicating strong and balanced classification. Conversely, low-resource languages like Lombard and Ligurian exhibit lower performance (0.85–0.87 F1-score). This disparity reflects the challenge of distinguishing closely related varieties and underscores the impact of training data quality, given the inherent noise present in the raw Wikipedia source texts. The complete performance metrics data can be seen in Table 2.

2.3 Data augmentation strategy

To address the severe parallel data scarcity inherent to our three target language pairs, we implemented a data augmentation strategy aimed at expanding our training set prior to model fine-

²<https://www.perseus.tufts.edu/hopper/>

Language	Precision	Recall	F1-Score
English	0.994	0.996	0.995
German	0.996	0.998	0.997
Latin	0.991	0.991	0.991
Italian	0.966	0.982	0.974
Piedmontese	0.998	0.984	0.991
Ladin	0.998	0.995	0.997
French	0.994	0.996	0.995
Spanish	0.995	0.996	0.995
Ligurian	0.968	0.954	0.961
Lombard	0.843	0.896	0.868
Romansh	0.945	0.903	0.923
Friulian	0.863	0.836	0.849

Table 2: Language detection performance metrics by language.

tuning. Our strategy relied on two complementary approaches: generating synthetic back-translation into English for all language pairs, and the incorporation of parallel data from related Romance languages for Ladin and Ligurian. The use of back-translated synthetic data aligns with established practices in low-resource MT settings (Sennrich et al., 2016; Ranathunga et al., 2023). Leveraging Large Language Models (LLMs) to generate synthetic parallel data via back-translation into a high-resource language (e.g., English) relies on the well-documented asymmetric capabilities of LLMs (Nguyen et al., 2024). Because LLMs exhibit significantly stronger generative fluency and semantic accuracy when mapping low-resource inputs into their English-dominant pre-training space, translating toward English yields high-quality target sentences that serve as reliable back-translated training data. We incorporated back-translated data into English for all three target languages: Icelandic, Ladin and Ligurian. The specific source monolingual corpora and the models selected for the back-translation process are detailed in the next section.

3 Training of the systems

3.1 English-Icelandic

3.1.1 Evaluation of existing models

In this subsection we present the evaluation of some of the existing models able to translate from English into Icelandic. This would help us choose which system to select for the fine-tuning process and to assess whether our fine-tuned model achieves competitive evaluation figures. In Table 3 we can observe the evaluation figures for the English-Icelandic language pair. The evaluation

was performed using the newtest2019 reference sets (eng-GB, for English) with 1,997 segments.

System	BLEU	chrF2	TER	COMET
SalamandraTA7b	29.7	56.1	57.2	0.8567
TowerPlus2B	32.8	58.6	56.6	0.8704
TowerPlus9B	38.3	62.9	49.4	0.8811
TowerPlus27B	39.5	63.3	49.2	0.8792
TranslateGemma4b	14.0	40.6	81.9	0.7050
TranslateGemma12b	20.3	50.6	72.1	0.8407
TranslateGemma27b	21.5	52.6	74.2	0.8591

Table 3: Evaluation results for some existing models for English-Icelandic.

Regarding the English-to-Icelandic translation task (Table 3), performance varies significantly depending on model family and parameter size. TowerPlus27B consistently achieves the highest scores across all surface-level metrics, peaking at 39.5 BLEU, 63.3 chrF2, and a low 49.2 TER, closely followed by its 9B counterpart, which yields a slightly higher COMET score of 0.8811. A clear scaling effect is evident within both the TowerPlus and TranslateGemma model families, as larger parameter counts systematically yield better translation quality. Notably, TranslateGemma models lag considerably behind in this direction, with the 4B variant underperforming substantially (14.0 BLEU, 81.9 TER), highlighting the inherent challenge of translating from English into a morphologically rich target language like Icelandic.

Conversely, we present the evaluation figures using the same reference corpus for the Icelandic-English language pairs (Table 4). These figures can guide us to select the model to perform back-translation of the monolingual Icelandic corpora.

System	BLEU	chrF2	TER	COMET
SalamandraTA7b	40.6	63.5	46.8	0.8768
TowerPlus2B	35.5	61.6	52.8	0.8708
TowerPlus9B	38.8	63.7	49.2	0.8787
TowerPlus27B	38.0	63.5	49.8	0.8773
TranslateGemma4b	24.6	53.5	67.2	0.8364
TranslateGemma12b	29.1	58.4	61.3	0.8600
TranslateGemma27b	30.1	59.7	58.8	0.8667

Table 4: Evaluation results for some existing models for Icelandic-English.

For the Icelandic-to-English language pair, performance across all systems shows a marked improvement, reflecting the reduced generation complexity when translating into English. Surprisingly, SalamandraTA7b outperforms all larger architectures in lexically oriented metrics, attaining the best overall BLEU (40.6) and TER (46.8). Meanwhile, TowerPlus9B demonstrates superior seman-

tic alignment, reaching top scores in chrF2 (63.7) and COMET (0.8787). Unlike the reverse direction, scaling parameter size within the TowerPlus family yields diminishing returns here, as TowerPlus27B performs slightly below the 9B model across all metrics. The TranslateGemma suite also exhibits substantial gains compared to the opposite direction, though it remains secondary to the leading systems.

Comparing both translation directions reveals a clear asymmetry in performance between translating into and out of Icelandic. All evaluated models achieve substantially higher baseline metrics when translating into English, primarily due to English having simpler inflectional morphology and broader representation in pre-training corpora. This contrast is particularly pronounced for SalamandraTA7b, which gains over 10 BLEU points in the Icelandic-to-English direction (40.6 vs. 29.7), effectively jumping from a mid-tier model into the top-performing position. Furthermore, while translating into Icelandic strongly favors scaled models like TowerPlus27B to handle complex target grammar, translating into English allows mid-sized models, such as TowerPlus9B and SalamandraTA7b, to compete with or even surpass their larger counterparts, making them optimal, computationally efficient candidates for back-translation pipelines.

Based on these evaluation results, SalamandraTA-7b-instruct³ was selected to perform the back-translation of the monolingual Icelandic corpora into English. Although TowerPlus9B marginally outperforms SalamandraTA-7b in chrF2 and COMET for the Icelandic-to-English task, this advantage is negligible compared to SalamandraTA-7b’s substantial lead in BLEU (+1.8 points) and TER (-2.4 points). Given its superior lexical accuracy, lower translation error rate, and greater computational efficiency, SalamandraTA-7b represents the most optimal candidate for generating high-quality synthetic parallel data. Furthermore, although it did not emerge as the top-performing model in the English-Icelandic evaluation figures, SalamandraTA-7b was selected for fine-tuning in accordance with our initial research design.

3.1.2 Corpora

To train and fine-tune the models we have used the parallel corpora and the back-translated corpora

³<https://huggingface.co/BSC-LT/salamandra-7b-instruct-2606>

shown in Table 5. We generated the back-translated segments from unaligned Icelandic Wikipedia segments (Table 1). These segments were first cleaned of cross-lingual noise using the `lid.176.bin` language detector, and subsequently back-translated using the SalamandraTA-7b-instruct. The MTUOC parallel corpus⁴ is created by concatenation, deduplication and rescoring of the NLLB, Paracrawl, HPLT and MacoCu corpora.

Corpus	Segments
Wikipedia	50,134
MTUOC	8,559,292
TOTAL PARALLEL	8,607,936
TOTAL BACK-TRANSLATED	244,689

Table 5: Size of the parallel corpora for English-Icelandic.

The training, validation and evaluation corpora have been created from the parallel corpus. Validation and evaluation sets have 5,000 segments.

3.1.3 Training Eole

To train the Eole models we used a configuration of an asymmetric Transformer sequence-to-sequence model—featuring a deep 24-layer encoder and a 6-layer decoder with a hidden dimension of 1024 and 16 attention heads—operating on a shared 64,000-token vocabulary. Leveraging multi-GPU distributed training in half-precision floating-point (fp16), it processed a batch size of 65,536 tokens per step (achieved via a 4,096-token per-device allocation and 4 gradient accumulation steps). The setup is optimized using Adam with a Noam decay schedule (peak learning rate of 1.5 and 8,000 warmup steps), complemented by standard regularization techniques like 0.1 dropout and 0.1 label smoothing. Finally, the training is capped at 200,000 steps, evaluating and checkpointing every 1,000 steps, while maintaining the 15 most recent checkpoints and applying early stopping after 10 non-improving validation rounds.

3.1.4 Fine-tuning SalamandraTA-7b-instruct

The training setup implements a full fine-tuning (FFT) procedure utilizing a ChatML format with specialized control tokens (`<lim_start>` and `<lim_end>`). Sequences are packed up to a maximum context length of 2,048 tokens and processed

⁴<https://huggingface.co/datasets/aoliverg/MTUOC-corpus-eng-ice>

with a global batch size of 256. Optimization is performed using the AdamW optimizer with a cosine learning rate scheduler, set to an initial learning rate of 2×10^{-5} , for a maximum of three epochs. To accelerate training and optimize memory bandwidth, the infrastructure employs bfloat16 precision, FlashAttention, and Fully Sharded Data Parallel (FSDP) across a multi-GPU environment. Model validation is conducted every 500 steps, integrating an early stopping mechanism based on evaluation loss (with a patience of three evaluation intervals) to automatically restore the best-performing checkpoint upon training completion.

The training and validation corpora have been converted into jsonl files with the following structure:

```
{
  "conversations": [
    {
      "from": "human",
      "value": "
        ↪ Translate the following text from English
        ↪ into Icelandic.\nEnglish: He actually
        ↪ received the award in year
        ↪ 1922.\nIcelandic:"
    },
    {
      "from": "gpt",
      "value": "
        ↪ Hann fékk verðlaunin 1922."
    }
  ]
}
```

3.1.5 Evaluation of the trained and fine-tuned models

As presented in Table 6, the evaluation of the three fine-tuning checkpoints of SalamandraTA on the English-to-Icelandic task reveals a notable divergence between lexical overlap metrics and neural semantic evaluation. Across all training checkpoints (cp10000, cp11000, and cp11500), model performance remains remarkably stable, indicating that training effectively converged by step 10,000 with minimal metric fluctuation thereafter. The fine-tuned SalamandraTA models consistently outperform the baseline NMT Eole model across all evaluated metrics. However, when comparing these values with the baseline evaluation of existing systems in Table 4, we observe that on this specific benchmark dataset (newstest2019), the fine-tuned models underperform relative to the base SalamandraTA-7b-instruct model.

System	BLEU	chrF2	TER	COMET
eval.en-Eole.is	24.1	51.4	61.6	0.7829
eval.en-cp10000.is	25.9	52.8	61.1	0.8245
eval.en-cp11000.is	25.4	52.3	62.2	0.8226
eval.en-cp11500.is	25.2	52.0	63.3	0.8217

Table 6: Evaluation results of the trained Eole and fine-tuned SalamandraTA-7b models for English-Icelandic.

After auditing the fine-tuning algorithms, datasets, and training pipeline without identifying

any operational errors, we re-evaluated both the trained Eole and fine-tuned SalamandraTA models on an in-house evaluation set. This dataset consists of 5,000 segments sampled from the parallel corpora described in Section 3.1.2. Table 7 details these results alongside the baseline performance of SalamandraTA-7b-instruct.

System	BLEU	chrF2	TER	COMET
SalamandraTA7b	37.0	59.2	52.1	0.8701
Eole	74.4	83.6	20.2	0.9219
cp10000	60.6	74.5	32.0	0.8988
cp11000	60.5	74.5	32.2	0.8978
cp11500	60.6	74.5	32.1	0.8973

Table 7: Evaluation results of the trained Eole and fine-tuned SalamandraTA-7b models for English-Icelandic with our evaluation set.

On this target evaluation set, both the Eole model and the fine-tuned SalamandraTA variants substantially outperform the base instruction-tuned model. We performed a strict overlap check across the training, validation, and test splits to rule out data contamination. Although these mixed results make it difficult to conclusively claim that the fine-tuned model outperforms the base model across all contexts, they demonstrate that both the training of Eole and the fine-tuning process of SalamandraTA functioned as intended. The marked performance gains on our evaluation set, along with the high metric stability across checkpoints, confirm that the models successfully adapted to the target dataset distribution. Consequently, checkpoint 11500 was selected for the final shared task submission.

3.2 English-Ligurian and English-Ladin

3.2.1 Existing MT systems

To our knowledge, there are not many machine translation systems available for these language pairs. The systems we identified and evaluated are: opus-mt-en-ROMANCE for Ligurian and Ladin, and GoogleTranslate for Ligurian. The performance of these systems is reported together with the results of our SalamandraTA-7b-instruct in subsection 3.2.4.

3.2.2 Corpora

As we detailed in Section 2.3, due to the extreme scarcity of parallel resources for English to Ligurian and Ladin, we expanded our data augmentation strategy beyond back-translation. Specifically, we combined back-translation into English,

executed using the gemma-4-31b-it model, with parallel data from closely related Romance languages. For these auxiliary Romance languages, we gathered both authentic parallel corpora and synthetic back-translated segments to maximize cross-lingual transfer, as shown in Table 8.

Language	Parallel	Back-translated
Ligurian	2,244	79,721
Ladin	4,340	455,775
Friulian	899	23,355
Lombard	5,196	51,373
Piedmontese	4,539	126,464
Romansh	1,268	8,956
Latin	15,733	340,392

Table 8: Number of parallel and back-translated segments per language.

These corpora derive from the Wikipedia extraction pipeline that we described in Section 2.1, supplemented by a subset of the WMT26 General MT task datasets. Specifically, for Ladin and Ligurian, we incorporated a selection of the parallel and monolingual resources provided by the task organizers. Since the parallel material was provided as unaligned monolingual files, we aligned and paired them into parallel corpora. Table 9 provides an overview of the task’s corpora that we used.

The validation and evaluation corpora have been extracted from the parallel corpora, with a size of 1,000 segments for Ladin and of 500 segments for Ligurian.

3.2.3 Fine-tuning SalamandraTA-7b-instruct

We used the same fine-tuning configuration described in subsection 3.1.4.

3.2.4 Evaluation of the fine-tuned models

System	BLEU	chrF2	TER
OpusMTROMANCE	2.2	28.9	91.2
GoogleT	24.0	49.9	64.5
FT-cp1000	16.4	44.2	76.3
FT-cp1500	18.1	45.7	73.9
FT-cp1911	18.2	45.6	73.6

Table 10: Evaluation for English-Ligurian.

We evaluated our full fine-tuned (FFT) systems across three distinct training checkpoints (cp1000, cp1500, and cp1911) and compared their performance against available baselines, specifically OpusMTROMANCE and Google Translate. Automatic evaluation was conducted using BLEU, chrF2, and TER metrics. It should be noted that

COMET scores are not reported, as the currently available COMET models do not provide support for these specific low-resource language pairs. The evaluation corpora are the one extracted from the real parallel corpora for Ligurian and Ladin as described in 3.2.2.

Table 10 presents the performance of the systems on the English–Ligurian task. While Google Translate achieves the highest performance across all metrics (24.0 BLEU, 49.9 chrF2), our FFT models demonstrate consistent progression as training advances. The model performance improves significantly from cp1000 (16.4 BLEU) to cp1500 (18.1 BLEU), reaching a plateau at cp1911 (18.2 BLEU). Although our fine-tuned systems fall short of Google Translate, they substantially outperform OpusMTROMANCE, which yields particularly poor translation quality (2.2 BLEU and a high TER of 91.2).

For English–Ladin, the evaluation results are reported in Table 11. Notably, Google Translate does not support this language pair. In this scenario, all our fine-tuned checkpoints outperform the OpusMTROMANCE baseline by a large margin. Our systems achieve optimal performance at cp1500 and cp1911, reaching up to 39.9 BLEU and 63.5 chrF2, while drastically reducing the TER from 98.3 (OpusMTROMANCE) to under 56. Overall, these results demonstrate the effectiveness of our fine-tuning approach.

System	BLEU	chrF2	TER
OpusMTROMANCE	25.4	37.4	98.3
FT-cp1000	36.8	61.9	62.0
FT-cp1500	39.9	63.5	55.8
FT-cp1911	38.9	63.5	55.7

Table 11: Evaluation for English-Ladin.

4 Participation in the shared task

We participated in the constrained track of the WMT2026 General Machine Translation Task for the following language pairs: English to-Icelandic, English-to-Ladin and English-to-Ligurian.

To perform the translation we started an MTUOC-server (Oliver, 2025) with the given model. Then we created a script that read the task’s json and translated the segments in the corresponding languages connecting to the MTUOC-server. To perform all the task, then, we used two fine-tuned SalamandraTA-7b-instruct models: one for English-Icelandic, and another for English to Ladin

Ligurian	Ladin
Monolingual Corpora	
Google-smol_sent-1-eng-lij_Latn	Sfrontull-la_usc_valbadia_loresmt24-1-lld
Consegioligure-linc-1-lij_Latn	Sfrontull-south_tyrol_weather_1ld-1-lld
Parallel Corpora	
Consegioligure-linc-1-lij_Latn	OPUS-tatoeba-v20230412-eng-1ld
	Motorola-lang_revitalization-20240604-eng-1ld
	OPUS-translatewiki-v20250101-eng-1ld
	OPUS-ubuntu-v14.10-eng-1ld
	OPUS-qed-v2.0a-eng-1ld

Table 9: Selected WMT26 datasets for Ligurian and Ladin.

and Ligurian.

5 Conclusion and Future Work

In this work, we presented the Grial-TA team’s submission to the WMT2026 General Machine Translation shared task (constrained track) for English into Icelandic, Ladin, and Ligurian. Our strategy relied on full fine-tuning (FFT) of the SalamandraTA-7b-instruct model, supported by a pipeline of custom parallel corpus extraction from Wikipedia, custom language identification, and LLM-based synthetic back-translation.

Our evaluation yields promising results across all targeted language pairs:

- **English–Icelandic:** While traditional sequence-to-sequence baselines (Eole) achieved higher surface-level lexical scores (BLEU), the fine-tuned SalamandraTA models demonstrated superior semantic adequacy and fluency, outperforming Eole in COMET (0.8190 vs. 0.7829).
- **English–Ladin & English–Ligurian:** For these extremely low-resource scenarios, cross-lingual transfer from related Romance languages combined with synthetic back-translation proved highly effective. Our fine-tuned models drastically outperformed available baselines (e.g., OpusMTROMANCE) on Ladin (reaching 39.9 BLEU), and built a competitive baseline for Ligurian (18.2 BLEU).

Although the official human and automatic evaluation rankings from the WMT26 Shared Task organizers are yet to be published, our internal benchmarks indicate that targeted full fine-tuning of mid-sized LLMs is a robust path forward for low-resource MT.

Open Science and Reproducibility: In commitment to open research, all code, processing scripts, and pipelines used in this work are publicly available on GitHub.⁵ Furthermore, the fine-tuned translation models can be directly downloaded and deployed as ready-to-use recipes for the MTUOC-server via Hugging Face.⁶

Future Work: Moving forward, within the framework of the *LLMTrad-IBE* project, we aim to extend the insights and methodologies established for Ladin and Ligurian to low-resource Romance languages spoken in the Iberian Peninsula, namely Asturian, Aragonese, Aranese, and Eonavian. We plan to explore refined multilingual fine-tuning paradigms, instruction tuning, and enhanced synthetic data generation techniques to further bridge the gap for lesser-resourced linguistic communities. Additionally, we plan to conduct an ablation study to systematically evaluate the real impact of our data augmentation strategies, in order to provide valuable insights for future research on low-resource languages.

Acknowledgments

This work has been supported by project "TaMTAS: Terminology-Aware Machine Translation for Accessible Science" - PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and by the European Union (CHIST-ERA).

This work is supported by the project "LLMTrad-IBE: Large Language Models for translating Low-Resource Languages of the Iberian Peninsula" granted by MCIU/AEI/10.13039/50110011033/FEDER,UE with reference PID2024-181570B-C33.

⁵<https://github.com/grial-research-group/WMT26-SharedTask-GeneralTranslation-GRIAL-TA>

⁶<https://huggingface.co/datasets/aoliverg/MTUOC-recipes>

This work is partially supported by the OpenEU Alliance, with funding from the European Union (EUROPEAN EDUCATION AND CULTURE EXECUTIVE AGENCY - EACEA) under grant agreement No 101177241.

This work was supported by the Industrial Doctorates Plan from the Department of Research and Universities of the Government of Catalonia (reference number: 2025 DI 00112).

References

- Jeff Johnson, Matthijs Douze, and Herve Jegou. 2021. [Billion-Scale Similarity Search with GPUs](#). *IEEE Transactions on Big Data*, 7(03):535–547.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. [Bag of tricks for efficient text classification](#). *Preprint*, arXiv:1607.01759.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Xuan-Phi Nguyen, 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: Long Papers)*, pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.
- Antoni Oliver. 2025. [MTUOC server: integrating several NMT and LLMs into professional translation workflows](#). In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 81–82.
- Antoni Oliver and Sergi Álvarez. 2023. [Filtering and rescoring the CCMATRIX corpus for Neural Machine Translation training](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 39–45, Tampere, Finland. European Association for Machine Translation.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- 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).
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Gemma Segué, Gonzalo López-Sánchez, and Antoni Oliver. 2026. Monolingual and Parallel Specialized Corpora from Wikipedia. In *Proceedings of the International Conference of the Spanish Society for Natural Language Processing (SEPLN)*, León, Spain. CEUR.org. Forthcoming.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- 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](#). *Preprint*, arXiv:2411.11295.
- Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. [Analysing Translation Artifacts: A Comparative Study of LLMs, NMTs, and Human Translations](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. Association for Computational Linguistics.

A Computing Infrastructure

All training and fine-tuning experiments were carried out on a high-performance computing node equipped with four NVIDIA RTX PRO 6000 Blackwell Edition GPUs (96 GB VRAM per card, totaling 384 GB VRAM). The hardware setup was driven by two 16-Core AMD EPYC 9135 processors (combining for 32 physical cores) alongside 384 GB of system RAM, executing on Ubuntu 22.04.5 LTS.

Appendix B shows the time and energy consumption of all the the training and finetuning processes performed during the shared task participation.

B Training and fine-tuning

B.1 Eole English-Icelandic

Total Time: 13h 34m 8s

Best Checkpoint (PPL): Step 75000
-> 9.16287 (ACC: 82.8373%)
Best Checkpoint (ACC): Step 86000
-> 82.9927\% (PPL: 9.22627)
Total GPU Consumption: 16.7649 kWh

We use the checkpoint 86000, corresponding to the best accuracy.

B.2 Eole Icelandic-English

Total Time: 11h 38m 35s
Best Checkpoint (PPL): Step 52000
-> 9.59479 (ACC: 80.1049%)
Best Checkpoint (ACC): Step 61000
-> 80.3643% (PPL: 9.60135)

Total GPU Consumption: 12.7221 kWh

We use the checkpoint 61000, corresponding to the best accuracy.

B.3 Fine-tuning SalamandraTA English-Icelandic

Total Time: 36h 44m 5s

Last 3 Checkpoints:
* Step 10000: PPL = 1.58200 | Loss = 0.45890
* Step 10500: PPL = 1.59400 | Loss = 0.46650
* Step 11000: PPL = 1.59800 | Loss = 0.46870

Total GPU Consumption: 48.0066 kWh

B.4 Fine-tuning SalamandraTA English-Ligurian and Ladin

Total Time: 6h 8m 9s

Last 3 Checkpoints:
* Step 500: PPL = 2.33800 | Loss = 0.84950
* Step 1000: PPL = 2.41200 | Loss = 0.88040
* Step 1500: PPL = 2.44200 | Loss = 0.89290

Total GPU Consumption: 7.9725 kWh