UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

Romance-MoLE: Rich Cousins and Their Benefits

An Approach to Low-Resource Language Modelling

Mihaela Goga

Essay/Thesis:	Master's thesis, 30 credits
Programme/Course:	Master in Language Technology, 120 credits
Level:	Second cycle
Semester:	Spring, 2026
Supervisor:	Jonas Lind, Anna Lokrantz and Asad Sayeed
Examiner:	Sharid Loáiciga
Keywords:	cross-lingual transfer learning, mixture of LoRA experts (MoLE), hierarchical mixture of LoRA experts (HMoRA), trans-tokenisation, occitan, parameter-efficient fine-tuning (PEFT)

Abstract

Integrating low-resource languages into Large Language Models (LLMs) is fundamentally limited by persistent data scarcity and the "Curse of Multilinguality," where dominant source languages overwrite minority language representations and cause structural "translationese". To preserve the endangered Languedocien dialect of Occitan without catastrophic forgetting, this thesis presents Romance-MoLE (Mixture of LoRA Experts), a parameter-efficient framework applying targeted cross-lingual transfer from Occitan's phylogenetic "rich cousins": Catalan and French. Using the pre-trained Llama-3.1-Carballo architecture, the pipeline applies Trans-Tokenisation to repair subword fragmentation and utilises curated synthetic data to enforce standardised dialectal norms. Independent Low-Rank Adaptation (LoRA) experts extracted for French and Catalan are fused via TIES-Merging to construct a stable baseline initialisation. Finally, a Hierarchical Mixture of Ranked Adapters routing strategy (HMoRA) applies token-level gating in shallow layers for localised morphological precision and sequence-level pooling in deep layers to eliminate mid-sentence code-switching. Romance-MoLE yields a +14.57 chrF++ improvement in Occitan translation quality over a standard fine-tuning baseline, while mitigating catastrophic forgetting in Catalan and allowing the model to follow French instructions with a significant reduction in structural code-switching.

Contents

1 Introduction

State-of-the-art Large Language Models (LLMs) have developed generative capabilities across domains at an unprecedented pace. However, this progress is unequally distributed, heavily biasing "Winner languages", that is, those with a dominant online presence, such as English (Joshi et al., 2020). Conversely, under-resourced languages like Occitan would fit the definition of "Scraping By", a status indicating that without organised, collective efforts to collect labelled data, they will be left behind by the AI progress. This lack of digital inclusion does not merely create a technological divide, it actively accelerates the digital erasure of the world's under-represented languages.

Occitan, a Western Romance dialect continuum historically spoken in Southern France, is particularly affected by this digital scarcity. Importantly, this scarcity leads to a lack of machine-readable text corpora, which is not just a technological side-effect, but the result of historical and systemic sociopolitical marginalisation (Blanchet, 2018). Occitan lacks official state recognition and has historically been wrongfully categorised by dominant institutions as a dialect, rather than a language. This distinction stems from a political rather than structural reason, since Occitan possesses its own fully developed grammatical, phonetic, and lexical systems derived directly from Latin (Harris and Vincent, 2003). Because it has historically been excluded from formal state governance, institutional education, and mainstream broadcast media, the language's formal digital footprint remains structurally constrained. While minority languages often foster active community participation in informal digital spaces such as social media, their lack of institutional data limits the availability of massive, machine-readable web corpora suitable for training datasets. Dialects like Languedocien, Provençal, Auvergnat and Limousin are currently classified by the UNESCO Atlas of the World's Languages (Moseley and Nicolas, 2010) as severely endangered.

While there has been a steady emergence of Occitan Natural Language Processing (NLP) research, such as Woller, Hangya, and Fraser (2021) exploring cross-lingual word embeddings, or Hopton and Aepli (2024) modelling its orthographic variations, it remains under-represented in the LLM era. Recent efforts have begun changing this, such as evaluations on Old Occitan part-of-speech tagging (Schöffel et al., 2025) and community-driven efforts to map the dialectal continuum (Nédey et al., 2026)

Despite localised efforts, integrating a low-resource language like Occitan into modern

LLMs presents several unresolved challenges. Brute-force pre-training requires initialising a model from scratch and exposing it to trillions of raw tokens to learn basic language structures. For Occitan, this is impossible, as its total available digital presence is significantly below the budget required to build a foundation model, which would serve as the general-purpose reasoning base. As a result, developing language technologies for Occitan forces the following dilemma: training a small, isolated monolingual model that lacks the advanced reasoning capabilities of modern LLMs, or relying on existing massive multilingual models.

To effectively preserve a low-resource language, relying on readily available multilingual models is not feasible as they suffer from the "Curse of Multilinguality" (Conneau et al., 2020). Because these models possess a fixed parameter capacity, attempting to represent too many disparate linguistic patterns causes catastrophic interference. In latent spaces, the internal layers, which encode both semantic concepts and structural rules, the statistical dominance of high-resource languages actively overwrites the representations of low-resource ones. Consequently, when prompted in Occitan, these models frequently route their reasoning through an English or French-aligned latent space, producing structural calques and *translationese* (Koppel and Ordan, 2011; Artetxe, Labaka, and Agirre, 2020). Furthermore, they indiscriminately blend distinct regional varieties, leading to the homogenisation of specific dialects like Languedocien (Bird, 2020; Kuparinen, Miletić, and Scherrer, 2023). The current literature lacks a parameter-efficient framework capable of bypassing this multilinguality curse, one that can enforce dialectal fidelity for Occitan without destroying the polyglot reasoning power of the base model.

1.0.1 Motivation

The motivation driving this thesis is deeply personal and contextualised within a wider ethical debate. As a speaker of Aromanian, a language classified as definitely endangered by UNESCO (Moseley and Nicolas, 2010) which shares a similar sociopolitical context as Occitan, I have witnessed this linguistic erosion first-hand. While my grandparents' generation spoke our language with significantly less majority language influence, the youngest generation suffers from heavier assimilation, due to the dual pressures of globalism and a lack of digital representation. Because our languages are systematically denied formal institutional support, they are victims of quiet, ongoing state assimilation.

Therefore, while the immediate scope of this research focuses on Languedocien Occitan, the underlying objective is creating a parameter-efficient, open-source framework that is reproducible and scalable on a standard consumer GPU. The long-term vision is that other researchers can improve and adapt it for other marginalised languages, giving them a chance to survive in the modern digital era.

1.0.2 Objectives

The main objective of this work is to bridge the knowledge gap by developing a parameter-efficient *polyglot* framework capable of modelling the Languedocien dialect of Occitan

alongside major high-resource languages. Rather than isolating Occitan in a monolingual space, this research proposes a targeted cross-lingual transfer strategy that actively combats the curse of multilinguality. By using Occitan's phylogenetic *rich cousins*: Catalan and French, the goal is to map constrained, dialect-specific data onto the pre-existing logic of a multilingual model, allowing the architecture to dynamically code-switch and translate between these languages without catastrophic forgetting, the abrupt loss of previously learned capabilities when new information is acquired.

The primary research question we seek to answer is: To what extent can phylogenetic initialisation (via TIES-Merging of French and Catalan adapters, task-specific modules added to a frozen base model) combined with dynamic routing via a Mixture-of-LoRA-Experts (MoLE) architecture, a mechanism that selects optimal expert pathways at the token or sequence level, mitigate the curse of multilinguality? Specifically, we investigate if this approach can prevent parameter interference and enable polyglot capabilities in both Occitan and its high-resource cousins.
The goal of this project is to provide community adaptability and the democratisation of language technology for endangered varieties. As an alternative to massive-compute solutions, this research targets standard consumer hardware, providing computational linguists, regional education institutions, and language preservation advocates with an open-source, scalable framework. By enabling an 8-billion-parameter polyglot model to preserve a minority dialect alongside dominant languages without performance degradation, it provides end-users with practical tools for localised translation, historical text digitisation, and authentic conversation modelling.

1.0.3 Methodology Overview

To answer this question, we utilise the Llama-3.1-Carballo model (Rodríguez et al., 2025), chosen for its strong pre-existing Ibero-Romance fine-tuning and latent French awareness. The methodology begins with a Trans-Tokenisation pipeline to improve the model's vocabulary and semantically initialise Occitan graphemes. Because Occitan data is heavily fragmented, a curated synthetic data pipeline is used to enforce Languedocien morphological norms.

With the foundation established, we extract individual Low-Rank Adaptation (LoRA) experts for French and Catalan and mathematically fuse them using TIES-Merging to create a stabilised initialisation. Finally, to enable dynamic token and sequence-level control during inference, we introduce the Romance-MoLE (Mixture of LoRA Experts) architecture, utilising a Hierarchical Mixture of LoRA Experts (HMoRA) strategy to balance the expert modules.

1.0.4 Contributions

The originality of this research does not stem from inventing a novel architecture from scratch. Rather, its main contribution lies in the application of state-of-the-art method-

ologies: Trans-Tokenisation, TIES-Merging, and HMoRA, to address the curse of multilinguality for a marginalised language.

The results show an improvement in polyglot dialect adaptation. The Romance-MoLE architecture outpaces the standard fine-tuning baseline, improving Occitan generative translation quality (chrF++) by over 14 points. The model isolates language representations by enforcing hierarchical routing. It recovers robust French and maintains Catalan capabilities, mitigating catastrophic mid-sentence code-switching, while ensuring coherent grammatical adherence to Languedocien Occitan.

However, our findings also suggest that while on one hand the languages mutually improve on each other via their phylogenetic similarity, on the other hand, that very same similarity acts as a double-edged sword. Because Catalan and Occitan are syntactically and morphologically close, their internal representation frequently overlap, causing the model to struggle with complex grammar differentiation. Furthermore, we observed that under token-level routing, French dominated the other languages and "translationese" remains a persistent challenge when working with cross-lingual transfer from a higher-resource to lower-resource language; though shifting to sequence-level routing significantly reduced this.

Properly differentiating between the languages demands a well calibrated and sensitive routing system, which is exceedingly hard. To achieve this, the model would require more diverse, contrastive and quality data (e.g., explicit code-switching examples that were absent in our training data). Ultimately, this still proves that specialised, multi-directional language preservation is possible within strict compute limitations by creatively combining existing frameworks.

In order to ensure transparency and experiment reproducibility, the Romance-MoLE framework alongside the evaluation scores have been made available here: `https://github.com/mihaelagoga/romance_mole`.

2 Background

Advancements in LLMs have yielded significant natural language capabilities, however, these gains remain disproportionately skewed towards the world's high-resource languages. Although modern architectures demonstrate robust zero-shot reasoning in dominant languages, their performance degrades dramatically when modelling low-resource and regional varieties (Robinson et al., 2023). Consequently, because the traditional method of scaling artificial intelligence relies on virtually endless data streams, adapting these massive models to severely under-represented languages like *Languedocien Occitan* requires a clear shift in methodology.

This chapter establishes the theoretical foundation for the research, examining the architectural patterns essential for adapting pre-trained LLMs to low-resource domains. It begins by exploring the inherent limitations of standard scaling, specifically the risks of cultural erasure and the negative interference known as the *Curse of Multilinguality* (Conneau et al., 2020; Chang et al., 2024). To overcome these barriers, the chapter reviews the mechanisms of cross-lingual transfer learning, detailing how to leverage phylogenetic "rich cousins" to bypass data scarcity. Finally, it explores the state-of-the-art solutions required to execute this transfer, including Trans-Tokenisation to resolve structural fragmentation, TIES-merging to prevent parameter interference, and hierarchical Mixture-of-Experts (MoE) routing to enable dynamic, multi-lingual inference without massive computational load.

2.1 A The Low-Resource Challenge in the LLM Era

2.1.1 The "Crossover Budget" and the Impossibility of Pre-training

The standard process for developing highly capable foundation models relies on pre-training from a random weight initialisation, a *blank slate* approach that requires an immense volume of data. High-resource models are typically exposed to trillions of tokens during their initial training phase to successfully internalise grammar and syntax, and to construct semantic representations of world knowledge. However, for low-resource languages, corpora of this magnitude simply do not exist (Touvron et al., 2023; Joshi et al., 2020).

The absolute limits of this data scarcity are mathematically formalised in the ATLAS

(Adaptive Transfer Scaling Laws) framework by Longpre et al. (2026). In their comprehensive analysis of multilingual scaling, the authors investigate the optimum compute boundaries of language modelling to identify critical "crossover points", the empirical thresholds that determine whether it is more effective to pre-train a model from scratch or to fine-tune an existing multilingual checkpoint. Their scaling laws demonstrate that pre-training from scratch only surpasses the performance of a fine-tuned model when the available training budget exceeds a massive threshold of 144 billion to 283 billion tokens, depending on the target language.

Applying these scaling laws to the Occitan context reveals a stark constraint. Because the total available digital footprint of the language is orders of magnitude below this 144-billion-token crossover budget, brute force pre-training for a modern, multi-billion parameter architecture is not only resource-intensive, but it is also not mathematically feasible. Since monolingual pre-training is not a viable option, achieving fluency in a low-resource language requires adopting a *polyglot* approach utilising cross-lingual transfer learning to map a highly constrained dataset onto the pre-existing logic of a massively multilingual model.

2.1.2 Prompting Failures and Translationese

Given the mathematical impossibility of pre-training an Occitan model from scratch, speakers have to rely on leveraging the multilingual capabilities of publicly available LLMs. However, while these architectures demonstrate strong zero-shot reasoning in dominant global languages, they consistently underperform in low-resource domains compared to specialised supervised baselines (Robinson et al., 2023).

This failure stems from the overwhelmingly English-centric design of publicly available foundation models. Recent interpretability studies reveal that when a massive LLM is prompted in a low-resource language, it does not think in that language by default (Wendler et al., 2024). Instead, the network frequently translates the prompt into an English-aligned latent space within its early layers, performs its logical reasoning in English, and translates the answer back into the target language in its final layers (Schut, Gal, and Farquhar, 2025). As demonstrated by Etxaniz et al. (2024), this internal translation bottleneck actively degrades the model's reasoning capabilities.

For a language such as Occitan, this internal translation bottleneck results in the generation of structural calques and "translationese", a phenomenon where translated outputs inherently adopt the statistical and syntactic signatures of the source language rather than native phrasing (Koppel and Ordan, 2011; Artetxe, Labaka, and Agirre, 2020). Rather than producing authentic, native-sounding outputs, the model generates sentences that use Occitan vocabulary but strictly adhere to English or French syntactic structures. Consequently, while the output might be technically comprehensible, it fails to capture the true grammatical mechanics and pragmatic flow of the language. This

means that relying on the zero-shot capabilities of generic commercial models is insufficient for high-quality language preservation, as the model's internal reasoning pathways must be explicitly aligned and structurally grounded in the target language.

2.1.3 Cultural Erasure and Dialectal Homogenisation

In addition to technical limitations, reliance on generic multilingual models for translation introduces a significant sociolinguistic risk: cultural erasure through dialectal homogenisation (Bird, 2020; Kuparinen, Miletić, and Scherrer, 2023). Standard NLP pipelines and LLMs operate under the assumption that all data within a language represent a unified monolith. However, Occitan is not a single uniform entity, instead it is a rich dialectal continuum that encompasses distinct regional varieties, such as Languedocien, Gascon, Provençal, and Limousin and it is crucially lacking a universally adopted orthography (Hopton and Aepli, 2024).

When modern foundation models are trained on unlabelled web-scraped corpora, they ingest these distinct, politically fragmented dialects indiscriminately. Because models lack dedicated metadata or dialectal training objectives, they fail to differentiate between regional morphologies and orthographies. This dialect blending results in the generation of a blended output, an unnatural mixture of dialects that does not exist in native speech. A readily available model will generate a sentence that improperly combines a Gascon definite article with a Provençal contraction and Languedocien verbal morphology. Consequently, these models entirely fail to adhere to standardised linguistic frameworks, such as the classical Languedocien norm (*Norma Classica*) established by Alibert (1935), and instead suffer from severe representation bleeding across closely related sub-variants (Aepli et al., 2023).

Furthermore, this predisposition to erasure is deeply rooted in the language's precarious sociopolitical reality. Historically subjected to state-sponsored linguistic assimilation and marginalisation in France, a systemic disenfranchisement often referred to as *la vergonha* ("the shame"), Occitan lacks a clear institutional and digital presence enjoyed by state languages (Blanchet, 2018).

To fully understand the sociolinguistic barriers facing Occitan NLP, its diglossic status must be explicitly stated. While the concept of diglossia was originally formalised by Ferguson (1959) to describe hierarchical variations within a single language, it was crucially expanded by Fishman (1967) to encompass societies where two entirely distinct languages occupy rigidly separated societal functions. In the context of Southern France, this manifests as a classic Fishmanian diglossia (Costa, 2023). French exclusively occupies the high prestige functions, dominating government, formal education, media, and digital infrastructure, while Occitan has been historically relegated to low prestige domestic and localised community interaction (Lafont, 1997). Because of this hierarchical separation, organic Occitan digital data is not only scarce but structurally subordinated, further emphasising why standard, uncurated NLP pipelines fail to capture the language in its true form.

Traditional NLP methodologies actively exacerbate this diglossia. To make raw text mathematically legible, standard preprocessing pipelines enforce "dialect-to-standard normalisation" (Kuparinen, Miletić, and Scherrer, 2023). By forcing Occitan text to conform to a single, often French-influenced computational standard, researchers inadvertently erase the unique morphological and phonetic traits of regional varieties (Blanchet, 2004). Ultimately, addressing the low-resource challenge requires more than merely generating comprehensible vocabulary: it necessitates an architectural approach capable of enforcing strict orthographic and dialectal boundaries to preserve the language's authentic cultural identity.

2.1.4 Overcoming Scarcity via Synthetic Data Generation

Digital scarcity is a common issue when it comes to low-resource languages, thus to mitigate it, various augmentation techniques such as synthetic data generation have been developed. Collecting human-labeled data is an arduous and time-consuming task which led to the usage of large foundation models to generate new text, taking on the role of generators for smaller, resource-constrained architectures (Pecher et al., 2026; Li et al., 2023). Pecher et al. (2026) demonstrated that even small amounts of synthetic data can enable smaller models to outperform larger models.

The main tenet behind this approach is the Knowledge Distillation (KD) technique. Originally, KD was designed to compress a large classifier into a smaller one by transferring logit probabilities from a *teacher* network to a *student* (Hinton, Vinyals, and Dean, 2015). Over time, it has evolved to be more commonly referred to as *Data Distillation* in the field of NLP, where a teacher model generates a custom training corpus (Wang et al., 2023). This way, student models can learn complex reasoning without the need for manual human annotation. Consequently, the field is increasingly refocusing on data quality as opposed to sheer volume. As shown by the LIMA study (Zhou et al., 2023) and the Phi models (Gunasekar et al., 2023), training on a smaller meticulously curated LLM-generated data supports better reasoning capabilities than unfiltered and noisy datasets. For instance, even current state-of-the-art models, like the Llama 3 family are reliant on synthetic data pipelines for their alignment and formatting (Grattafiori et al., 2024).

While synthetic data generation has its advantages, it is not a flawless replacement for human data. Relying solely on it increases the risk of embedding structural calques and propagating hallucinations. Since student models learn to imitate the stylistic fluency of their teacher without the associated grounding, they often generate out-of-distribution vocabulary that looks stylistically good but lacks genuine semantic meaning (Gudibande et al., 2023).

2.2 Cross-Lingual Transfer Learning and Phylogenetic Scaffold

2.2.1 The Strategy of Weight Initialisation

As formalised by the ATLAS crossover framework (Longpre et al., 2026), pre-training a foundation model from a random weight initialisation is a mathematical impossibility for Languedocien Occitan due to severe data scarcity. In a random initialisation, a neural network starts with arbitrary parameter values, meaning it starts as a blank slate with no understanding of syntax, grammar, or vocabulary. To bypass this hard constraint, modern NLP relies on Cross-Lingual Transfer Learning (CLTL) (Conneau et al., 2020; Chang et al., 2024), a strategy that shifts the computational burden away from the target language and onto pre-existing, high-resource architectures.

During the pre-training phase of massive multilingual models, exposure to trillions of tokens allows the architecture to construct a highly complex, multidimensional latent space (Touvron et al., 2023). Within this space, the model encodes abstract, universal linguistic representations, learning fundamental mechanics such as syntactic dependencies, semantic clustering, and logical reasoning pathways (Wendler et al., 2024). Cross-lingual transfer exploits this pre-existing architecture. Instead of forcing a model to learn the fundamental concept of language from scratch using a highly constrained Occitan corpus, the model is initialised with these previously optimised parameter weights.

The primary objective of this warm-start methodology is to harness the established semantic understanding of the base model. Using these robust initialisation weights, the time and data required for the fine-tuning process are drastically reduced. The model is no longer tasked with learning universal grammar, rather, its objective is narrowed to mapping the specific lexical items and surface-level morphological rules of the Languedocien dialect onto an already functioning, globally aware latent space. However, while cross-lingual transfer addresses the problem of sheer data volume, the efficiency of this mapping is highly dependent on the structural compatibility of the initialisation weights (Woller, Hangya, and Fraser, 2021), meaning that transfer learning cannot be treated as a language-agnostic process (Chang et al., 2024).

2.2.2 The "Rich Cousin" Hypothesis and Topological Proximity

Although cross-lingual transfer learning provides a mechanism to bypass data scarcity, it remains highly susceptible to negative interference if the source and target languages are improperly matched. The current literature identifies a significant limitation in massively multilingual models (such as mBERT or XLM-R) known as the "curse of multilinguality" (Wang, Lipton, and Tsvetkov, 2020; Conneau et al., 2020). As the number of languages in a fixed-capacity model increases, performance on individual languages degrades because the network's parameters are forced to represent too many disparate linguistic patterns.

To mitigate this, recent research advocates for a targeted regional strategy, which restricts the transfer pool to a small cluster of phylogenetically related languages (Chang et al., 2024). According to the Adaptive Transfer Scaling Laws (ATLAS) framework, the absolute strongest predictor of positive transfer is sharing a language family and script, which statistically outweighs the benefits of sheer model scale. This validates the selection of a Romance-centric base model such as Llama-3.1-Carballo (Rodríguez et al., 2025), which is extensively pre-trained on Galician, Catalan, and Portuguese over a generic English-centric architecture.

Crucially, the mechanics of this transfer are predominantly driven by syntax rather than vocabulary. Chang et al. (2024) demonstrated that syntactic similarity (shared grammatical structures, such as Subject-Verb-Object word order) accounts for the vast majority of performance improvements in cross-lingual transfer, whereas lexical overlap provides only marginal gains. Therefore, imperfect vocabulary alignment between Occitan and its high-resource neighbours is an acceptable limitation, provided the syntactic scaffolding remains intact.

For Languedocien Occitan, French and Catalan serve as the optimal phylogenetic "rich cousins." By transferring parameters from these specific languages, we can implement a complementary merging strategy as demonstrated by Kunz, Debess, and Simonsen (2025) in their work on Faroese. Merging adapters from distinct but related source languages creates an initialisation that possesses a hybrid of strengths. By combining Catalan (which shares morphological and phonetic roots with Occitan) and French (which provides semantic reasoning capabilities), the model inherits grammatical fidelity and logical reasoning from the first token.

2.2.3 Bridging the Orthographic Gap: Trans-Tokenisation

Even when leveraging optimal phylogenetic cousins, low-resource transfer is frequently bottlenecked at the embedding layer. Standard multilingual tokenisers are heavily biassed towards high-resource languages and treat non-standardised or dialectal text as a sequence of fragmented subwords or raw byte fallback tokens. Because these fallback tokens rarely appear in consistent, meaningful contexts during pre-training, their corresponding embedding weights remain poorly optimised and carry minimal semantic value (Remy et al., 2024). For Occitan, this orthographic fragmentation wastes critical context space and forces the model to rely on less meaningful semantic representations, severely limiting its downstream performance (Remy et al., 2024).

Traditionally, replacing a model's tokeniser requires re-initialising the embedding table from scratch, which causes a substantial degradation in pre-trained performance (Remy et al., 2024). To resolve this without destroying the foundational knowledge of the model, this research employs Cross-Lingual Vocabulary Transfer, specifically through the framework of Trans-Tokenisation (Remy et al., 2024; Sakajo et al., 2025).

Rather than relying solely on orthographic character matching, Trans-Tokenisation uses translation resources (such as bilingual dictionaries or aligned parallel corpora) to map semantic meaning across languages (Remy et al., 2024; Sakajo et al., 2025). Using a probabilistic alignment strategy, the embedding for a new Languedocien token is initialised as a weighted average of its semantically similar source tokens from French and Catalan. For example, the model does not need to relearn the abstract concept of "beautiful", it simply learns that the Occitan token *polida* maps directly to the pre-existing conceptual weights of its Romance counterparts. This semantic mapping effectively bypasses the orthographic variations of the dialect, allowing the Occitan adapter to inherit the latent structural knowledge of its rich cousins while preserving the original hidden layers of the LLM.

2.3 Parameter Synthesis and Hierarchical Routing

Another challenge consists in integrating multi-lingual knowledge without causing catastrophic interference, a well-documented phenomenon where learning new tasks overwrites previously acquired network parameters (McCloskey and Cohen, 1989; Kirkpatrick et al., 2017). In the context of Large Language Models, standard transfer pipelines typically degrade the foundational logic of the base model, precisely because of this forgetting. To preserve the zero-shot reasoning of the base LLM, while dynamically injecting domain-specific linguistic knowledge, recent advancements in parameter-efficient fine-tuning focus on a hybrid architecture utilising TIES-Merging and HMoRA.

2.3.1 Adapter Fusion via TIES-Merging

In parameter-efficient transfer learning (PETL), specialisation is often achieved by means of independent modular weight updates known as adapters (Houlsby et al., 2019). Unlike standard fine-tuning, which requires training a completely new model for every task, adapter modules provide an extensible alternative by adding only a fraction of trainable parameters while the base model's parameters remain fixed (Houlsby et al., 2019). To integrate multi-domain knowledge without suffering catastrophic forgetting to sequential training, the researchers introduced Adapter Fusion (Pfeiffer et al., 2021). This method attempts to avoid the downsides of parameter interference by splitting the learning process into distinct phases. A knowledge extraction phase, where individual adapters are trained in isolation to capture specific linguistic or task-based behaviours, followed by a knowledge composition step, which dynamically merges the modules. By separating the acquisition of new skills from their integration, the framework enables the base model to acquire diverse capabilities in a non-destructive manner (Pfeiffer et al., 2021).

The current state of adapters includes Task-Arithmetic (Ilharco et al., 2023), a framework demonstrating that neural network behaviour can be realigned by treating fine-tuned weight updates as "task vectors" capable of being manipulated via linear algebra. By subtracting the weights of the pre-trained base model from a fine-tuned checkpoint, we extract isolated language vectors.

However, simple element-wise addition of these vectors often leads to negative parameter interference. When merging closely related domains, models frequently exhibit "sign conflicts", where parameter updates move in opposite mathematical directions, causing shared representations to neutralise each other (Yadav et al., 2023). To resolve this, TIES-Merging is employed (Yadav et al., 2023). This technique isolates the core features by Trimming redundant parameters, electing a dominant sign direction to resolve conflicts, and executing a Disjoint Merge. This mathematically ensures that the resulting initialisation retains the strongest complementary features of its parents' domains without parameter cancellation, providing a superior starting point compared to random initialisation.

2.3.2 Dynamic Mixing: The X-LoRA Framework

The foundation for compressing these adaptations is the standard Low-Rank Adaptation (Hu et al., 2022), a framework originally developed to address the prohibitive computational costs of full-parameter fine-tuning on limited hardware. Rather than updating an entire model, LoRA freezes the pre-trained base weights and injects trainable low-rank decomposition matrices directly into the transformer layers. This ensures strict memory efficiency by allowing the model to adapt while updating only a fraction of the total parameters.

However, the standard LoRA relies on a single set of learnt matrices, restricting the model to a static state. Although techniques like TIES-Merging provide a highly optimised initialisation by mathematically combining these matrices, this static merging ultimately forces the model to rely on a single averaged representation. To process highly complex or domain-shifting inputs—such as natural code-switching in multilingual environments, models require dynamic, token-level control rather than static averages.

To achieve this combinatorial flexibility, modern architectures employ the Mixture of Low-Rank Adapter Experts (X-LoRA) framework (Buehler and Buehler, 2024). Following the precedent of upcycling pre-trained dense models into sparse Mixture-of-Experts architectures (Komatsuzaki et al., 2023), the X-LoRA paradigm overcomes the limitations of static merging by maintaining multiple distinct LoRA experts simultaneously. In

this architecture, the base model and the individual domain adapters remain completely frozen. Instead, a lightweight gating network is introduced at each layer. This router dynamically evaluates the hidden states of the model and assigns probability weights to the distinct experts at the token level (Buehler and Buehler, 2024). Consequently, the router can selectively activate one specific expert to process a localised feature (such as a domain-specific named entity) while immediately shifting probability weights to other adapters to handle the surrounding syntactic structure.

2.3.3 Hierarchical Routing and Auxiliary Loss

A limitation of standard MoE implementations is the use of uniform routing strategies across all layers, which ignores the structural granularity of Large Language Models. Research shows that shallow layers primarily capture fine-grained lexical and token-level nuances, while deeper layers process broader semantic and task-level understanding (Liao et al., 2025).

To optimise this linguistic processing pipeline, modern frameworks employ the Hierarchical Mixture of LoRA Experts methodology (Liao et al., 2025). By implementing a structural layer threshold, these models use token-level routing in the shallow layers to rapidly navigate localised, word-by-word morphological differences. Consequently, in the deeper layers, the architecture transitions to sequence-level routing, often achieved via mean-pooling of the hidden states—to enforce global grammatical and logical consistency across the broader context window.

The introduction of dynamic routing carries the risk of "expert collapse," a common failure state where the gating network over-relies on high-resource domain adapters while ignoring low-resource or highly specialised experts. To prevent this monopolisation, an auxiliary load balance loss (L_{bal}) is typically injected during training. Modelled after Switch-Transformer penalties (Fedus, Zoph, and Shazeer, 2022), this mathematical constraint penalises routing imbalance and forces genuine specialisation among the available modules, ensuring that the model remains anchored to its targeted training objectives.

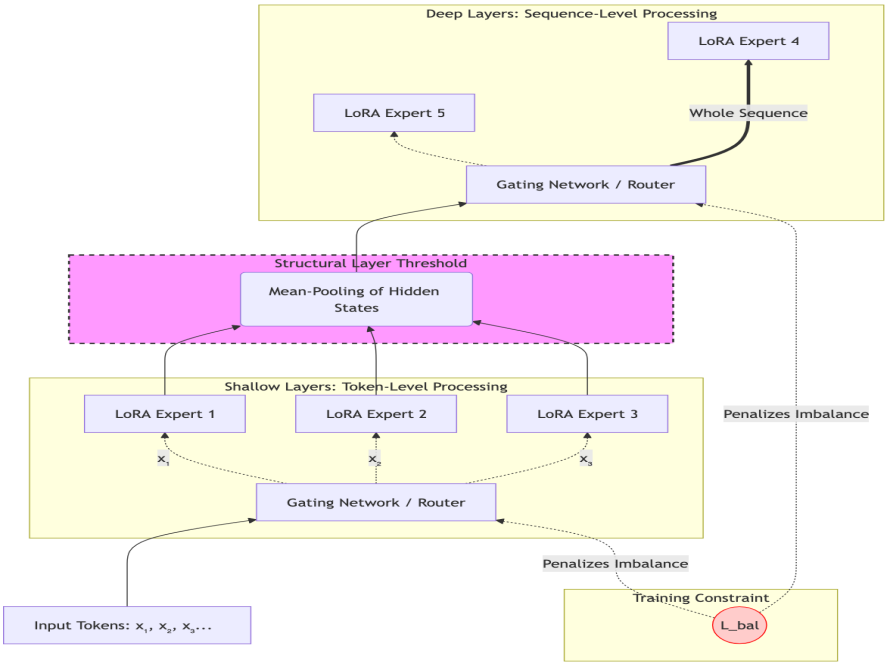

Figure 2.1: The Hierarchical Mixture of LoRA Experts (HMoRA) architecture.

3 Method

3.1 HPLT Data

As the backbone for the Occitan training, we used the HPLT-V3 Occitan dataset provided by the HPLT Consortium (Burchell et al., 2025). The dataset is a mixture of Common Crawl[1] (92.03%), the Internet Archive[2] (7.97%) and curated data of generally Occitan origin. As stated by the authors, the dataset has not been cleaned. Therefore, we performed extensive data cleansing to improve its quality for use in the Occitan Expert Curriculum Training.

The clean-up proceeded as follows: First, we applied regular expression filters to discard Wikipedia edit tags, such as *[modificar la font]*, and other web-scraping boilerplate text. We identified these elements by analysing the frequency of n-grams in the data. Next, we implemented length constraints to discard documents with fewer than five words. This step removed auxiliary items, such as titles, navigation elements, and other lexical noise, which lack grammatical value. To ensure that the remaining data was exclusively Occitan, we removed any sentences lacking the official HPLT language identification tag (Appendix Subsection A.3.1). Subsequently, we shuffled and split the corpus using a fixed deterministic seed (42) to prevent data leakage and support curriculum learning. Finally, we reserved 1000 entries for the development set, 2000 for the test set, and the remainder, 71837, for the training set.

After cleaning, the dataset was reduced from approximately 75 million tokens to only about 48 million tokens. We prioritised linguistic purity, this aggressive filtering improved the data quality. As highlighted by Remy et al. (2024) and Joshi et al. (2020), the massive multi-billion-token corpora required for standard, full-parameter LLM training simply do not exist for low-resource languages. Furthermore, our methodology was strictly constrained by hardware limitations. The training environment was restricted to two Nvidia A30 GPUs, each providing 24GB of VRAM. Standard full-parameter fine-tuning of large foundation models requires extensive memory to store optimiser states and gradients, which vastly exceeds this 48GB total VRAM budget. This dual constraint of limited data and limited compute required a shift towards Parameter-Efficient

[1] `https://commoncrawl.org/`
[2] `https://archive.org/`

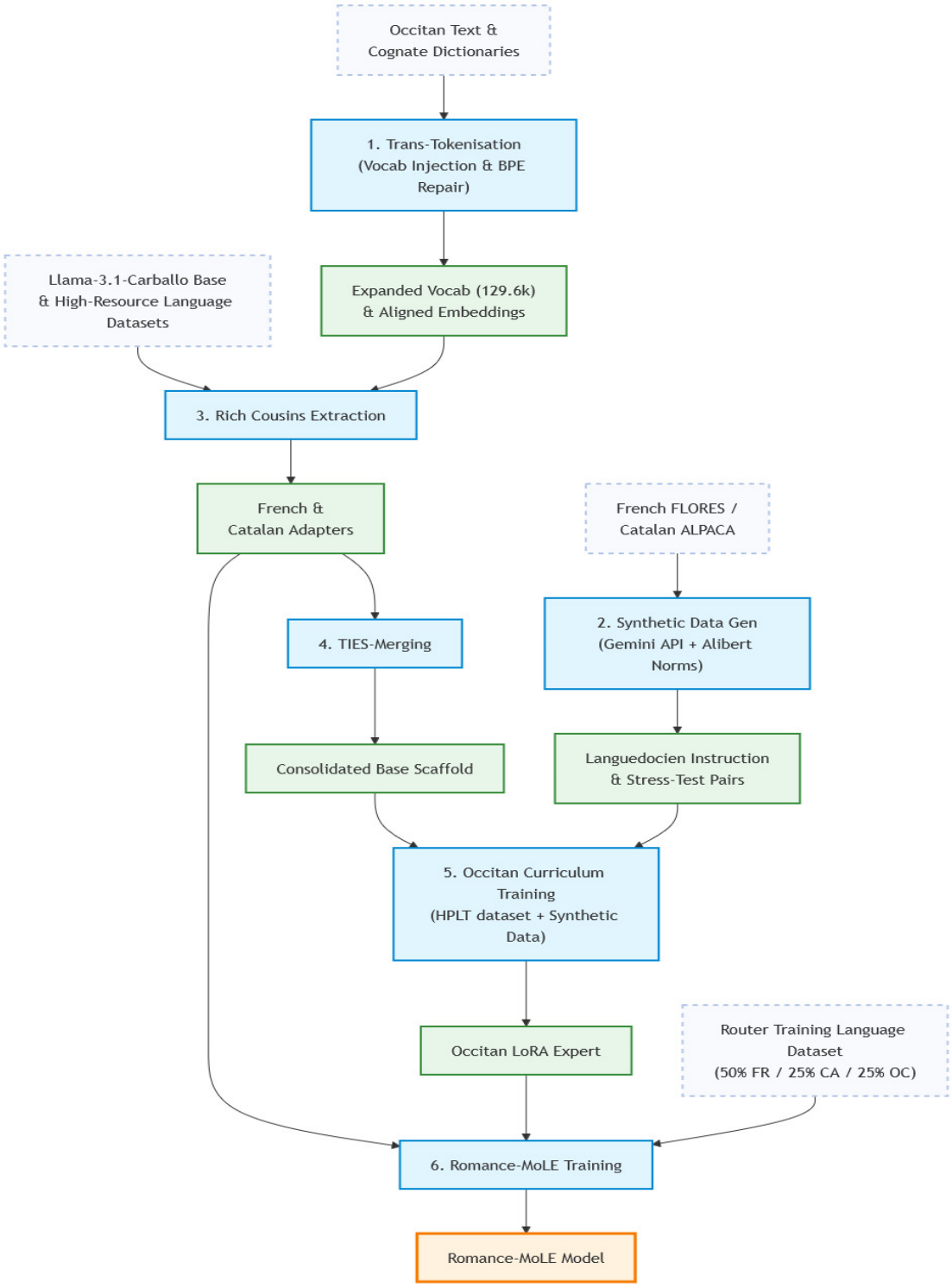

Figure 3.1: General overview of the Romance-MoLE training pipeline

Fine-Tuning (PEFT) techniques, specifically Low-Rank Adaptation (Hu et al., 2022). By freezing the weights of the base model and updating only a small number of injected low-rank matrices, LoRA significantly reduces VRAM consumption. This approach allows the model to achieve performance comparable to full fine-tuning while updating only a fraction of the parameters (Hu et al., 2022; Zadouri et al., 2024), making it the necessary structural choice for adapting low-resource languages like Occitan within a strict hardware budget.

3.2 Synthetic Data

To overcome the data scarcity problem established in the previous section and facilitate the development of the Occitan expert model, we created a high-quality, synthetic dataset in the Languedocien dialect of Occitan. We specifically chose Languedocien as the foundational basis for the *Norma Classica* orthography (as discussed in Section 2.1.3) to ensure a baseline for computational modelling. This pipeline for synthetic data generation was implemented specifically to address the lack of dialect-specific data, since the foundational HPLT-V3 corpus contains unlabelled and unstructured text encompassing mixed regional variants (Languedocien, Gascon, Provençal, and Limousin).

This constrained high-fidelity synthetic data generation was directly inspired by the LIMA framework (*Less Is More for Alignment*) (Zhou et al., 2023). LIMA's Superficial Alignment Hypothesis demonstrates that models acquire their core capabilities during pre-training; therefore, alignment does not require massive datasets, but rather a limited number of high-quality and stylistically varied examples. Following this methodology, we prioritised structural diversity and strict grammatical adherence to enforce the Languedocien identity over sheer data volume.

We used the Gemini API (Google DeepMind, 2025b; Google DeepMind, 2025a) for synthetic data generation, heavily constrained by system prompts (Appendix Chapter A) to enforce the standardised orthography of the Languedocien dialect. The selection was motivated by the baseline multilingual capabilities of the Gemini Pro and Flash models, as demonstrated by the state-of-the-art multilingual reasoning performance and zero-shot grammatical adherence when evaluated on the Multilingual Multidiscipline Language Understanding (MMMLU) benchmark (Google DeepMind, 2025b; Google DeepMind, 2025a). Furthermore, to ensure they provided the necessary transfer learning foundation to bridge high-resource Romance cousins to low-resource Occitan, we validated their cross-lingual translation capabilities ourselves using the FLORES-200 (Costa-jussà et al., 2022) evaluation suite.

To optimise both performance and computational budget, the generation was split across two models with their reasoning parameters maximised: Gemini 3 Pro Preview was utilised for generating the complex Languedocien feature-sets FLORES-200 (Costa-jussà et al., 2022) localisations, morpho-stress datasets meant to enforce targeted morphological features, and multi-turn dialogue exchanges, while Gemini 3 Flash Preview was

deployed for the higher-volume Alpaca-style (Taori et al., 2023) instruction dataset due to its similar linguistic performance and higher rate limits.

Inspired by the principles of Expert Prompting (Xu et al., 2023) and role adherence (Wang et al., 2024a), Gemini was specifically instructed to act as a linguist and expert translator specialising in the Languedocien dialect, adhering strictly to the classical Norm (*Norma Classica*) established by Louis Alibert (Alibert, 1935).
The comprehensive system prompt (provided in full in Appendix Chapter A) enforced grammatical rules including:

- **Morphological Elisions:** Strict enforcement of pre-vowel elisions (e.g., de → d', lo → l').
- **Dialectal Containment:** Explicit suppression of Gascon articles (eth, era) and Provençal contractions (dau, pòu), forcing the Languedocien equivalents (lo, la, del, pel).
- **Orthographic Standardisation:** Strict adherence to Occitan-specific graphemes (e.g., lh, nh, ò, ç) and the enforcement of straight typographical apostrophes (').
- **Verbal Morphology:** Alignment with Alibert's official conjugation tables.

To strengthen the model's tokeniser boundaries against the heavy fragmentation typical of low-resource languages, a specialised "Morphological Stress" dataset was synthetically generated. Comprising 250 sentences, this dataset concentrated statistical edge-cases designed to explicitly test and fine-tune subword tokenisation limits. The generation was strictly categorised into three domains: *Elision Overload* (100 sentences forcing 3-4 mandatory pre-vowel elisions per sentence, such as *qu'es* and *m'agrada*, to test boundary splitting), *Diacritic Density* (100 sentences heavily featuring specific Occitan open vowels and cedillas like *ò, à, è,* and *ç* to enforce UTF-8 byte mapping logic), and *Dialect Shibboleths* (50 sentences enforcing hyper-local Languedocien contractions while strictly penalising Gascon or Catalan equivalents to prevent cross-dialect token bleeding).

Because native internet scrapes, such as the HPLT corpus, overwhelmingly skew toward formal or written registers, a supplementary "Spoken Dialogue" dataset was synthesised to capture authentic conversational pragmatics. This dataset consisted of 250 multi-turn exchanges spanning 25 mundane, real-world scenarios (e.g., haggling at a market or ordering at a café). However, accurately representing colloquial Occitan presents a distinct challenge due to the dialect's historical lack of a uniformly adopted orthography in informal contexts. By enforcing the strict Norma Classica on colloquial structures, we introduced a necessary but artificial blend of spoken pragmatics and formal written complexity. To encourage the outputs to feel naturally spoken rather than rigidly translated within these constraints, the generation prompt targeted a high density of clitic pronouns (e.g., *m', t', li, ne, i*) and complex stacked combinations (e.g., *Te'n vas?* or *Dóna-m'o*). Natural interrogatives, colloquial idioms, and strict negative constraints

against French structural calques were additionally applied, ensuring the model could fluidly process highly complex conversational subword targets.

Although an Occitan FLORES-200 benchmark exists, it suffers from the same dialectal mixture as the HPLT corpus; using it would introduce Gascon or Provençal features, risking cross-dialectal contamination. Therefore, sentences were instead localised from the French FLORES-200 training dataset. Within modern localisation theory, standard translation often only transfers meaning in a narrow semantic sense, failing to align models with a specific community. As Hershcovich et al. (2022) argue, the "assumption that there is a cross-lingual, cross-culturally common semantics to preserve fails when the common grounding does not match between cultures." Adapting the French dataset required genuine localisation, targeting the specific Languedocien 'locale' (Pym, 2004) by modifying the source text to reflect regional sociocultural norms, toponyms, and pragmatic expectations, rather than performing a generic, one-to-one linguistic transfer.

To verify that our synthetic dataset was genuinely independent and not just French syntax masked with Occitan vocabulary, we conducted a comprehensive post-hoc validation. We aligned the 500 generated French-to-Occitan FLORES sentences directly against their native Occitan reference translations. Comparing these pairs across both surface-level and semantic metrics revealed zero exact string matches. The resulting synthetic outputs showed only a 54.48% character 5-gram overlap and a mean normalised Levenshtein distance of 0.3377. Together, these scores confirm that the model produced distinct surface forms rather than simply copying the source structure. At the same time, the outputs successfully preserved the intended meaning, achieving a strong chrF++ score of 59.13 and a mean sentence-embedding cosine similarity of 0.8343. Finally, the synthetic subset displayed an authentic level of lexical diversity, closely matching the native reference baseline's type-token ratio (0.3886 vs. 0.3971).

While the localised FLORES-200 data grounded the model semantically, it was necessary to transition the base model from a generic next-token predictor into an instruction-following assistant. To achieve this, synthetic instruction tuning was implemented using a cleaned Alpaca dataset. A total of 2,000 instruction-response pairs were extracted from Catalan, chosen due to its phylogenetic closeness to Occitan. Irrelevant or non-local topics (e.g., US laws, corporate entities) were removed via regex filtering. A custom system prompt (Appendix Chapter A) then instructed Gemini to replace non-Occitan geography, companies, and civic structures with Southern French/Occitan equivalents, establishing a native cultural baseline. Furthermore, strict rule-following constraints (Wang et al., 2024a) were applied to dissociate the model from its assistant identity, ensuring that the model did not refer to itself as an AI or translate out of Occitan, forcing it to act strictly as a Languedocien-centric assistant.

To test our synthetic generation pipeline, a native Occitan pedagogical expert [3] quali-

[3] Jossiane Mote, personal communication, April 2026. Qualitative review of Languedocien datasets.

tatively evaluated a representative sample of the generated datasets encompassing both multi-turn dialogues and formal text. The results confirmed the framework is overall structurally sound: the expert concluded that the underlying syntax was "globalement oui" [generally correct] and noted that the model successfully applied complex morphological conjugations conforming to Alibert (1935)'s standard. In addition, the generated texts were generally intelligible, proving that the superficial alignment constraints established a functional Occitan structure.

Despite these foundational successes, the qualitative analysis revealed limitations inherent in cross-lingual transfer from high-resource Romance languages. The expert identified that while the model's morphology was accurate, its stylistic phrasing occasionally defaulted to French or Catalan calques. For instance, the model exhibited an over-reliance on the compound past tense (*passé composé*) instead of the authentic Occitan *preterit* for completed past actions, and occasionally mapped Catalan lexical structures (e.g., using the verb *téner* for possession instead of the correct Occitan *aver*). Instances of lexical hallucination were observed when the model attempted to translate specific, out-of-distribution concepts. For example, when prompted to generate descriptive nature vocabulary, a domain likely underrepresented in the foundational Occitan corpus, the model hallucinated non-existent verbs such as *s'endetha* (intended as 'melts') and *pôdre* (intended as 'to bloom'), making those specific sentences incomprehensible to the native speaker.

In practice, the scarcity of native Languedocien experts and the time constraints of the project meant this qualitative review had to function as a final diagnostic rather than an ongoing feedback loop. While Ranathunga et al. (2023) argue that regional communities should ideally take the lead in dataset creation to ensure cultural context and accurate judgement of the bias, acting as a single researcher made the societal-scale participation described by Nekoto et al. (2020) impossible for the scope of the project. While the expert identified these stylistic and temporal biases, we could not retroactively update the generation prompts with more complex syntactic rules, such as those for the Occitan *preterit*. This reflects the fundamental barrier in language modelling: the continuous, human-driven alignment that drives such models (Ouyang et al., 2022) would be exceedingly difficult to implement for endangered languages.

3.3 Base Model Selection and Phylogenetic Scaffolding

The foundational architecture chosen for the Occitan Expert model is Llama-3.1-Carballo-8B (Rodríguez et al., 2025). The selection of this specific model over other open-source alternatives was motivated by a strategy of cross-lingual transfer learning, specifically leveraging the phylogenetic proximity of Occitan to high-resource "rich cousin" languages: Catalan and French.

Training an 8-billion-parameter language model for a low-resource language from scratch is computationally prohibitive and practically impossible due to data scarcity. Conse-

quently, it was necessary to adopt a pre-trained architecture and adapt it. However, adapting a model that lacks foundational Romance language comprehension yields sub-optimal cross-lingual alignment. Therefore, selecting an architecture with pre-existing Romance language exposure was paramount. The Llama-3.1-Carballo model provides an ideal linguistic scaffold for two critical reasons:

- **Native Catalan Fluency:** The Carballo variant was explicitly fine-tuned on regional Iberian languages, giving it high-quality syntactic and semantic representations of Catalan. Because Catalan and Languedocien Occitan share a tight phylogenetic and morphological overlap, the model's existing attention heads require minimal weight updates to align with Occitan grammar.
- **Latent French Representations:** While Carballo specialises in Iberian languages, it inherits the massive multilingual pre-training of the base Meta Llama 3.1 architecture. Consequently, the model possesses robust, latent knowledge of French.

By utilising the Carballo model, the objective of the training pipeline shifts from teaching the model a new language in a vacuum, to exposing and bridging its existing French and Catalan neural sub-networks. This phylogenetic scaffolding drastically reduces the computational budget required for convergence, as the model can rely on its pre-existing high-resource weights to infer the semantic meaning of low-resource Occitan inputs.

3.4 Trans Tokenisation

Even with high-quality data, standard Large Language Models struggle with low-resource languages due to tokenisation inefficiencies (Remy et al., 2024; Sakajo et al., 2025). Because the Carballo model inherits the stock Llama 3.1 tokeniser, it suffers from high token fertility when processing Occitan, splitting words into excessive, meaningless sub-word fragments. This fragmentation degrades the model's morphological understanding and artificially shrinks its effective context window (Remy et al., 2024). To address this, a four-step Trans-Tokenisation pipeline was implemented to safely patch and expand the existing tokeniser without destroying the pre-trained weights of the base model.

3.4.1 Targeted Vocabulary Injection and Byte-Level Repair

The pipeline targeted morphological expansion. Rather than guessing required tokens, Occitan elisions (e.g., *d'aquò*, *l'òme*, *qu'es*) and highly frequent diacritic words (e.g., *çò*, *tèrra*, *bòria*) were selected. Simultaneously, a corpus-mining algorithm scanned the training data to extract frequently fragmented words, utilising a filtering threshold (minimum frequency of 5, minimum length of 3) to prevent vocabulary pollution.

A frequent risk when adapting LLMs to new languages involves the HuggingFace fast-tokeniser mishandling specialised graphemes (Ahia et al., 2023; Petrov et al., 2023). Because Llama-3 utilises Bytelevel Byte-Pair Encoding (BPE) (Grattafiori et al., 2024),

injecting raw strings directly corrupts the decoding tables. To prevent this, the Occitan characters were explicitly encoded in their BPE Unicode string representations prior to addition. Single diacritics (e.g., *à* or *è*) were blocked from being added as tokens.

Furthermore, a tokeniser repair script was executed to resolve vocabulary ID collisions. Any added token configuration mapping to an original base vocabulary ID (ID $< 128{,}256$) was stripped. This prevented the fast-tokeniser from coercing existing subwords into raw text, eliminating the `U+FFFD` decoding corruption artefacts that commonly affect low-resource tokeniser patches.

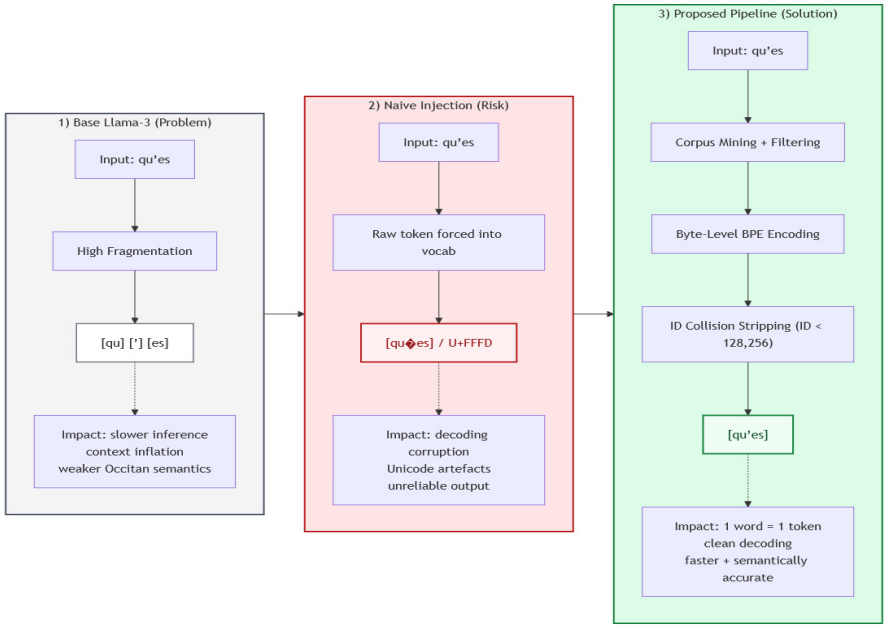

Figure 3.2: Comparison of tokenisation strategies. (1) Baseline tokeniser heavily fragments Occitan words. (2) Naive injection of raw text leads to decoding artifacts. (3) BPE injection and collision repair

3.4.2 Semantic Initialisation via Cognate Alignment

While the tokeniser can parse Occitan with greater efficiency, the newly added tokens lacked semantic representation in the model's embedding space, naively initialising with random, untrained weights. To solve this, the pipeline mathematically bridged the gap between Occitan and its close linguistic relatives (Remy et al., 2024; Sakajo et al., 2025).

Using Gemini (`gemini-3-pro-preview` with thinking budget set to high) as a cross-lingual mapper, the newly injected Occitan tokens were translated into their French and Catalan equivalents, with prompt constraints (Appendix Subsection A.2) to preserve

morphological structures identical to the Occitan root. To determine how strongly the Occitan token aligned with its French and Catalan cousins, a structural similarity weight (λ) was formulated. This was achieved by computing the cosine similarity of character bigrams ($n = 2$) between the Occitan word and its translations. A smoothing epsilon ($\epsilon = 0.01$) was added before normalising the similarities to alignment ratios.

During embedding initialisation, if a French or Catalan cognate was fragmented into multiple base-model tokens, the corresponding token embeddings were averaged to obtain a single representation. The source vectors were not L2-normalised before interpolation, since both French and Catalan cognate vectors were drawn from the same pretrained embedding table and already shared the same learned coordinate system and magnitude scale. Instead, the new Occitan embedding vector was initialised as a normalised weighted average of the available representations:

$$v_{oci} = \frac{\lambda_{fr} v_{fr} + \lambda_{ca} v_{ca}}{\lambda_{fr} + \lambda_{ca}} \tag{3.1}$$

Where λ_{fr} and λ_{ca} are the alignment weights derived from the French and Catalan cognate evidence. This averaging step prevents uncontrolled growth in vector magnitude while preserving the base model's original embedding scale. (*Note: Tokens that failed translation fell back to random initialisation, tracked via a generated metadata log.*)

Finally, to ensure computational efficiency during the subsequent training phases, the expanded vocabulary was purposefully padded to a final size of exactly 129,600 tokens. This specific dimension is a multiple of 64, which heavily optimises memory alignment and throughput on NVIDIA CUDA tensor cores.

3.4.3 Analysis of Sub-Optimal Fragmentation

The resulting Trans-Tokenisation yielded improvements that were marginal in absolute scale but highly visible and structurally significant. While the patched tokeniser reduced complete clitics like *qu'ei* to a single token, instances like *l'òme* demonstrated a partial reduction (from 4 tokens to 3). Rather than a failure of ByteLevel mapping, this highlights the necessary balance between reducing token fertility and preventing vocabulary bloat.

In the case of *l'òme*, the tokeniser merged the elided article into a cohesive token ([*l'*]), isolating it from the root noun ([*òme*]), which retained some byte-level fragmentation. From a linguistic perspective, this partial reduction is highly advantageous: it preserves the boundary between the grammatical article and the semantic root. Forcing every elided noun into a single monolithic token would rapidly exhaust the 129,600 vocabulary limit and result in sparse, poorly trained embeddings. Ultimately, reducing the sequence from 4 to 3 tokens still represents a 25% computational efficiency gain per instance, significantly expanding the model's effective context window without corrupting its underlying morphological awareness.

3.5 Rich Cousins LoRAs

To utilise the phylogenetic scaffolding established by the base model selection, it was necessary to isolate and extract the model's native understanding of Occitan's closest high-resource relatives: Catalan and French. Rather than updating the full parameters of the base model, which risks catastrophic forgetting of its generalised linguistics representations, Low-Rank Adaptation (LoRA) was utilised. Two distinct "Rich Cousin"s expert adapters were trained on top of the Occitan-initialised, trans-tokenised base model.

3.5.1 The Catalan Pipeline: Exploiting Iberian Alignment

Because the Llama-3.1-Carballo base architecture was already heavily fine-tuned for Iberian regional languages, extracting the Catalan expert required an instruction-tuning pipeline. A targeted dataset of 41,600 instruction-response pairs was extracted from the `saillab/alpaca-catalan-cleaned` (Upadhayay and Behzadan, 2024) corpus. Data were formatted using a standardised Alpaca (Taori et al., 2023) prompt template to elicit clear follow-up behaviour. Due to the model's strong pre-existing Ibero-Romance alignment, this volume of data was sufficient to saturate the Catalan adapter weights.

3.5.2 The French Pipeline: Counteracting Base-Model Bias

In contrast, extracting the French expert presented a significant architectural challenge. Although the model retained latent French knowledge from its foundational Meta Llama 3.1 pre-training, its recent Carballo fine-tuning instilled a heavy Ibero-Romance bias. To counteract this bias and steer the model towards becoming a Gallo-Romance expert, the French data pipeline was intentionally scaled up and aggressively augmented.

- **Volume Scaling and Augmentation:** The initial extraction targeted 110,000 examples from the `jpacifico/French-Alpaca-dataset-Instruct-110K` (Pacifico, 2024) corpus, double the volume of the Catalan baseline. This was further increased by merging and deduplicating multiple secondary French instruction tuning datasets (e.g., `Alpaca_french_instruct` (Boukhari, 2023), `Alpaca_french_mixtral` (AIffl, 2024)). An automated filtering protocol removed `<unk>` token artefacts and unmapped inputs to ensure dataset quality.
- **High-Fidelity Repair Set:** To complement the rigid nature of standard instruction datasets, a specialised conversational "repair set" was curated. High-fidelity, human-annotated dialogue trees strictly filtered for positive review scores and non-deletion statuses were extracted from the `OpenAssistant/oasst2` (Köpf et al., 2023) dataset. This ensured that the French expert captured authentic conversational pragmatics.

3.5.3 Adapter Architecture and Training Dynamics

Both experts were trained using the `SFTTrainer` framework with highly specific Low-Rank Adaptation targets (Hu et al., 2022). To maximise learning capacity while main-

taining parameter efficiency, the LoRA rank was set to r=64 with a scaling alpha of $\alpha = 128$.

The adaptation matrices were injected deeply into the model's architecture, targeting the attention mechanisms (*q_proj*, *k_proj*, *v_proj*, *o_proj*), the MLP projections (*gate_proj*, *up_proj*, *down_proj*), and crucially, the language modelling head (*lm_head*). Modifying the *lm_head* natively permitted output tokenisation matching the newly patched characters in the target languages. However, the *embed_tokens* layer was strictly frozen to prevent inference-time mismatches with the custom Trans-Tokenisation dictionary established in (Chapter 3.4). While untying the input and output layers preserved cross-lingual semantics (Press and Wolf, 2017; Inan, Khosravi, and Socher, 2016), it introduced the risk of logit drift, where the trained output diverges from the frozen input space, risking artificial overconfidence in new subwords. To counteract this, a stabilisation function continuously monitored the cosine distance between the flattened *embed_tokens* and the *lm_head* matrices. If this drift exceeded a 10% threshold (cosine distance > 0.1), the matrices were automatically averaged to enforce an explicit embedding tie (Press and Wolf, 2017). This realigned the output distribution with the base vocabulary, preventing significant drift without restricting the initial learning phase.

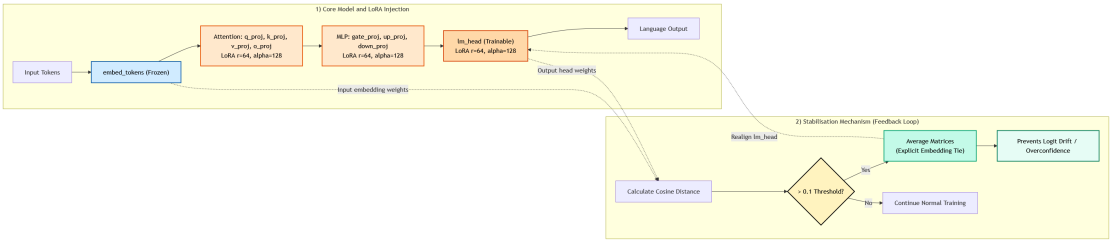

Figure 3.3: LoRA adapter architecture and stabilisation dynamics

To optimise training throughput across the 1024-token context window, the sequences were concatenated using `packing=True`. Because simple Alpaca dialogue responses contain significant padding (e.g., hundreds of dead tokens, tokens which do not contribute to the model's understanding), which force the GPU to perform empty calculations, sequence packing chains multiple separate instructions sequentially to entirely remove variable pad-token waste. However, utilising packed sequences with standard PyTorch Scaled Dot Product Attention (SDPA) inherently risks "cross-contamination," where attention maps bleed context into logically separate examples sharing a sequence block. To safely execute this, *FlashAttention-2* (Dao, 2023) was strictly enforced, since it correctly applies boundary-aware diagonal blocking natively, yielding training operations along with massive attention map VRAM savings.

Because the baseline Llama-3.1 8B model requires approximately 16 GB of VRAM strictly to exist in `bfloat16` precision, executing highly dense instruction datasets within tight 24 GB GPU hardware bounds required memory-efficiency protocols:

- **Paged AdamW:** As introduced by Dettmers et al. (2023), it was used to prevent hard CUDA Out-of-Memory (OOM) crashes during backward pass accumulations, the memory manager was configured to seamlessly page standard AdamW optimiser momentum and variance metrics to system CPU RAM during memory spikes.
- **Gradient Checkpointing:** Instead of storing the massive computation graph of activations required for backpropagation, the GPU deleted intermediate layers and recalculated them on the fly during the backward pass (Chen et al., 2016). This swapped a computational time penalty for a critical $\sim$ 70% activation memory saving.
- **Microbatching and Accumulation:** To avoid erratic batch-normalisation convergence induced by small batch sizes, the pipeline utilised strict microbatching (`per_device_train_batch_size=1`) while calculating and holding mathematical gradients over 16 sequential steps (`gradient_accumulation_steps=16`). This simulated a robust effective batch size without blowing past the active memory envelope.

The training loops executed over 2 epochs with an automated 10% validation split, using an `EarlyStoppingCallback` (Wolf et al., 2020) monitored against evaluation loss to preserve the optimal checkpoint.

Finally, to maximise the impact of French conversational data, an explicit oversampling strategy was deployed during the French adapter training. The high-quality `oasst2` repair data was artificially injected at a 3x higher relative rate than the augmented synthetic data, effectively forcing the adapter to prioritise conversational fluency over generic instruction-following, aligning with the Superficial Alignment Hypothesis (Zhou et al., 2023).

3.6 TIES-Merging of Rich Cousin Adapters

With French and Catalan experts trained through LoRA adaptation, the next architectural requirement was to fuse these independent knowledge representations into a singular "Occitan Initialisation" adapter. Simply averaging the weights of two distinct adapters typically results in catastrophic parameter interference and degraded performance. To mitigate this and reduce theoretical hallucination risks, the TIES-Merging (Trim, Elect, Sign, and Merge) methodology introduced by Yadav et al. (2023) was implemented using the `mergekit` framework.

3.6.1 The 50/50 Linguistic Continuum Balancing

The most important hyperparameter in the merging process was the relative weighting of the experts, since an unbalanced ratio would bias the resulting weights too much in

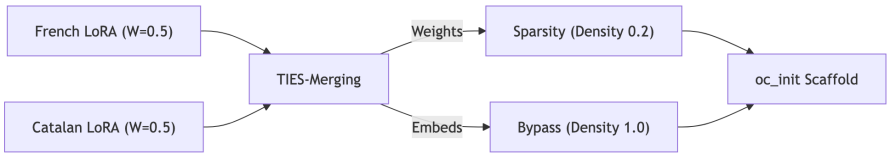

Figure 3.4: TIES-merging process

favour of one particular language. To ensure equal linguistic influence, we implemented a perfectly balanced split (`WEIGHT_FR=0.5`, `WEIGHT_CA=0.5`).

While Languedocien Occitan is phylogenetically and syntactically closer to Catalan (Chang et al., 2024), it shares a significant modern lexical overlap with French. However, assigning a higher weight to the Catalan adapter was deemed unnecessary and potentially detrimental. Because the underlying Llama-3.1-Carballo base model possesses a massive pre-trained Ibero-Romance bias (Rodríguez et al., 2025), this "Catalan pull" remains latently active even within the French LoRA weights. Consequently, a strict 50/50 parameter merge organically resolves to a $> 50\%$ Catalan inferential bias during generation. This anchors the merged adapter to Occitan's true position on the language continuum without requiring manual scaling, mirroring the complementary merging strategies observed in cross-lingual initialisations (Kunz, Debess, and Simonsen, 2025).

3.6.2 Density Sparsification and Vocabulary Protection

To combine the adapters, the TIES-Merging algorithm (Yadav et al., 2023) computes the parameter deltas, filters them according to sign agreement (consensus masking), and applies sparsification. To minimise the risk of hallucinatory degradation during cross-lingual transfer, the density threshold was chosen empirically and was heavily constrained (`DENSITY=0.2`), retaining only the top 20% of parameter magnitudes from each adapter (Yadav et al., 2023; Kunz, Debess, and Simonsen, 2025). The consensus mask was cast into `int8` space for computational efficiency, and the deltas were dynamically normalised to maintain magnitude scale.

However, a key architectural exception was integrated into the sparsification logic. Any tensors interacting with the expanded vocabulary, specifically `embed_tokens` and the `lm_head` were completely bypassed (`density=1.0`). Subjecting the language modelling head to 20% sparsification would have instantly corrupted the custom Trans-Tokenisation matrices (Chapter 3.3), destroying the model's ability to decode the newly injected Occitan graphemes.

3.6.3 Post-Merge Stabilisation and Embedding Tying

Since embedding representations can drift independently during isolated adapter training, a post-merge stabilisation protocol was executed to ensure geometric coherence. The

pipeline computed the cosine distance (drift) between the newly merged `embed_tokens` and the `lm_head` vectors.

A strict threshold of 0.1 (10% drift) was established. If the computed drift surpassed this limit, the pipeline programmatically enforced an "embedding tie." As established in the foundational language modelling architecture (Press and Wolf, 2017; Inan, Khosravi, and Socher, 2016), this was achieved by mathematically averaging the divergent matrices into a single tensor and cloning the result back into both the embedding and head slots, guaranteeing representational stability.

Finally, to prevent initialisation failures when loading the merged `PeftModel`, the state dictionaries were key-cleaned. Legacy PEFT artefacts, such as `.modules_to_save` and `.original_module` suffixes, were stripped to resolve key collisions, and the adapter configurations were passed through a strict metadata whitelist before final export into the `adapter_oc_init` directory.

3.7 Occitan Curriculum Training and Base Consolidation

Following the TIES-merging of the French and Catalan adapters, the resulting `adapter_oc_init` baseline provided a robust scaffold. However, because this initialisation does not actually possess any explicit Occitan language knowledge, it was necessary to execute a highly constrained fine-tuning process. To impose a pure Languedocien Occitan ability, a two-stage sequential curriculum training strategy was implemented.

3.7.1 Architectural Overrides and Prompt Anchoring

Standard Low-Rank Adaptation (LoRA) conventionally freezes the input and output language embeddings to optimise GPU memory. However, to successfully overcome the parent language biases and properly align the trans-tokenised vocabulary established in Section 3.3, this default behaviour was overridden. The training script dynamically resised the base model and explicitly assigned both `embed_tokens` and the `lm_head` to the PEFT `modules_to_save` configuration. This architectural override unlocked the matrices, allowing the model's fundamental vocabulary representations to update natively during gradient descent.

To prevent the model from drifting or code-switching back into French or Catalan during training, a "System Prompt Anchor" was used (Wang et al., 2024a). When processing conversational datasets, the tokeniser pipeline automatically overrode the prompt head with a strict, localised directive: *"Sès un assistent d'intelligéncia artificiala que parla unicament en occitan lengadocian..."* (You are an artificial intelligence assistant that speaks strictly in Languedocien Occitan). This explicit prefix anchored the model's internal attention states to a pure Languedocien output mode.

In addition, a data collator was deployed to process batches. It automatically injected -100 label masking across the user prompt tokens, ensuring that backpropagation and

loss calculations were strictly limited to the assistant's generated responses (Taori et al., 2023). It also reduced ID ranges against maximum vocabulary targets to prevent `IndexOutOfRange` crashes during experimental sub-word tokenisation.

3.7.2 The Two-Stage Curriculum

The curriculum was designed to sequentially transition the model from foundational language acquisition to complex pragmatic alignment.

- **Stage 1: Foundational Linguistic Adaptation:** The first phase utilised the heavily filtered HPLT corpus (detailed in Section 3.1). The objective of this pre-training phase was to direct the model towards general Occitan mechanics, grammar structures, and common idioms. The sequence length was constrained to 512 tokens with a learning rate of 5×10^{-5} and 0.0 weight decay.
- **Stage 2: Dialect & Instruction Subspace:** Following up on the Stage 1 checkpoint, the second phase used exclusively the synthetic dataset. The sequence length was expanded to 1024 tokens to accommodate complex conversational structures. Crucially, to prevent catastrophic overfitting on the relatively small synthetic dataset, NEFTune (Noisy Embedding Instruction Fine-Tuning) (Jain et al., 2023) was applied. By injecting random uniform noise ($\alpha = 1$) directly into the embedding vectors during the forward pass, NEFTune acted as a powerful regulariser, forcing the model to generalise its instructional alignment rather than memorising exact synthetic phrasing. Overfitting was further mitigated via an `EarlyStoppingCallback` dynamically monitoring the validation loss gap.

3.8 Romance-MoLE: Hierarchical Architecture and Router Training

The final architectural phase of the project required embedding the newly formed Occitan expert (Section 3.7) with the previously extracted French and Catalan "Rich Cousin" LoRA adapters (Section 3.4). Rather than utilising standard Parameter-Efficient Fine-Tuning (PEFT) adapter swapping, which limits the model to a single language space at a time, a dynamically routed architecture was applied: the Romance-MoLE.

3.8.1 The Core Routing Equation and HMoRA Strategy

The broad concept of Romance-MoLE is to dynamically utilise the outputs of multiple frozen LoRA language experts on the fly. During the forward pass, the native computation shifts into a sum-product algorithm governed by trainable router matrices, which compute the gate weights via a softmax over all available experts:

$$y = W_{base}(x) + \sum (\texttt{gate\_weights}_i \cdot B_i(A_i(x)) \cdot \texttt{scaling}_i)$$

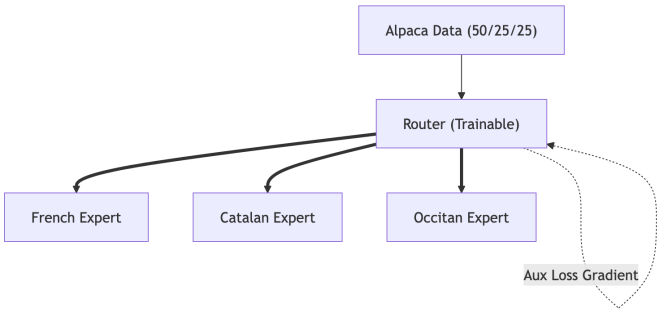

Figure 3.5: Romance-MoLE overview

While the underlying adapters remain strictly frozen, the linear routing components (`nn.Linear`) dynamically compute the optimal parameter combinations based on the input context.

However, standard token-level routing frequently results in structural code-switching, where a model oscillates between languages mid-sentence. To resolve this, a depth-sensitive Hierarchical Model Routing (HMoRA) strategy (Liao et al., 2025) was adapted, using a `sequence_route_threshold` of 8:

- **Shallow Layers (Layers 0-7, Token-Level Routing):** The early layers of the network generate independent expert gates per individual token. This grants the model the granular flexibility required for low-level morphological and syntactic adjustments.
- **Deep Layers (Layers 8+, Sequence-Level Routing):** At or above the 8th layer, the architecture mean-pools its hidden states iteratively. Consequently, all tokens within a given sequence receive an identical set of `gate_weights`. This forces the model to make coherent, higher-order language decisions, effectively eliminating structural code-switching across paragraphs.

3.8.2 Router Curriculum and Data Preparation

To ensure that the router accurately mapped language representations, a custom Alpaca-formatted curriculum was constructed. The data set was structured using an asymmetric and deliberate linguistic ratio: 50% French, 25% Catalan, and 25% Occitan.

The heavy weighting of French data (sourced from `timpearce/alpaca-cleaned-french` (Pearce, 2023)) was a deliberate choice to balance out the base model's inherent Ibero-Romance biases, ensuring the router did not disproportionately collapse toward Catalan representations. The Occitan data consisted of the merged splits stage (Subsection 3.7.2), and the Catalan data was sampled from `saillab/alpaca-catalan-cleaned` (Upadhayay and Behzadan, 2024). Crucially, the preprocessing script rigorously tagged

each instruction with its parent language code, establishing a robust basis for implicit `expert_target` mapping during training.

3.8.3 Auxiliary Loss and Autograd Injection

A known vulnerability of routing models is "expert collapse" (Wang et al., 2024b), where the router monopolises toward the initially strongest expert and starves the others. To mitigate this, a Switch-Transformer Load-Balancing auxiliary penalty was formulated (Fedus, Zoph, and Shazeer, 2022) ($L_{bal} = E \cdot \sum(f_i)^2$) and applied using a `router_aux_loss_coef` of 0.005.

Implementing this mathematically in a distributed environment presented a significant challenge. Advanced parallelism frameworks, such as Fully Sharded Data Parallel (FSDP) and gradient checkpointing, inherently sever the global accumulations of auxiliary losses (Zhao, Gu, Vujanic, et al., 2023). To bypass this framework limitation, a modified PyTorch autograd module (`InjectAuxLoss`) was used to execute a targeted gradient injection (Ganin et al., 2016) (implementation in Appendix Chapter B). The underlying class boilerplate was adapted from the open-source implementation by van Vugt (2018). By overriding `torch.autograd.Function`, the module silently patched the load-balancing penalty gradients directly into the backward pass at the specific layer level, preserving complete structural integrity for the internal optimisers without breaking the computation graph.

3.8.4 Distributed Data Parallel (DDP) and Sparse Checkpointing

To strictly adhere to hardware VRAM constraints (dual 24GB VRAM GPUs), the MoLE training pipeline inherited the foundational memory optimisation protocols established during the Rich Cousin LoRA extraction. Furthermore, sequence lengths were tightly capped at 128 tokens, as the router logic efficiently evaluates short conversational contexts without incurring the quadratic memory overhead of full-document attention.

Building upon this heavily optimised baseline, a targeted Multi-GPU framework was required. Since active router networks consist of a highly compact parameter footprint (∼400k parameters), utilising heavy-sharding frameworks like ZeRO (Rajbhandari et al., 2020) or FSDP introduced unnecessary overhead and socket redundancy bugs (e.g., duplicate NCCL buffer creations causing Out-Of-Memory crashes (Zhao, Gu, Vujanic, et al., 2023)). Therefore, the pipeline was built using PyTorch's native `DistributedDataParallel` (DDP). To ensure strict GPU isolation and prevent IPC collisions, the script pre-emptively spliced `CUDA_VISIBLE_DEVICES`, sandboxing each concurrent process to perceive only a single physical GPU at boot. This effectively bypassed the catastrophic inter-GPU peer-to-peer (P2P) buffer penalty.

To accelerate convergence, explicit supervision hooks were integrated. These hooks applied a Cross-Entropy penalty (`router_supervision_coef`) early in the training phase

against the tagged language codes, forcing the network to explicitly map specific languages to their respective adapters before the generative loss fully stabilised the weights. Finally, a sparse checkpointing protocol was implemented. The `MoLEDDPTrainer` overrode the default `.save()` method, deliberately skipping the duplication of the base model weights and sub-LoRAs. The pipeline strictly isolated and saved the `router_weights.pt` and `router_config.json`, resulting in a highly modular, deployment-ready Romance-MoLE architecture.

4 Results

This chapter presents the empirical evaluation of the Occitan training pipeline and the Romance-MoLE architecture. The MoLE router was trained with a sequence routing threshold of $t = 8$. This threshold means that transformer layers with index < 8 use token-level routing, allowing local lexical and subword-sensitive expert variation, while layers with index ≥ 8 use sequence-level routing to ensure a stable global expert choice throughout generation. It has a load-balancing auxiliary loss ($\lambda_{\text{aux}} = 0.01$) to discourage expert collapse, while adding explicit language-label router supervision ($\lambda_{\text{sup}} = 0.02$) for monolingual examples. The evaluation is divided into three primary experiments: translation quality and distributional alignment (Section 4.1), the recovery of multilingual capabilities via dynamic routing (Section 4.2), and the preservation of fine-grained grammatical features (Section 4.3). A qualitative analysis of the routing heatmaps follows to showcase the hierarchical gating mechanics.

4.1 Generation Quality and the Alignment Tax

This first evaluation compares the multi-stage curriculum and Trans-Tokenisation pipeline (`FullPipeline`) against a single-stage LoRA fine-tune of the base model on the HPLT corpus (`Simple`). The models were evaluated on held-out FLORES-200 translation task sentences (French to Occitan and Catalan to Occitan). To capture surface-level features, while remaining sensitive to the complex morphology and minor orthographic variations of Occitan, we evaluated target generation using chrF++ character n-gram overlap. Furthermore, because standard token-level perplexity is incomparable across different tokenisers, we evaluate intrinsic model likelihood using conditional Bits-Per-Character (cBPC), providing a fair, information-normalised measure of exact-sequence probability. Statistical significance for all translation quality metrics was determined using paired bootstrap resampling (Koehn, 2004), a standard non-parametric test for machine translation evaluation.

As shown in Table 4.1, the curriculum pipeline outperforms the simple baseline in generative translation across both source languages.
On the `fr` $\rightarrow$ `oc` translation, the full pipeline scored 45.45 compared to 30.88 (+14.57, 95% CI [13.73, 15.41], $p = 0.002$). This advantage persists in the `ca` $\rightarrow$ `oc` direction (+11.58, 95% CI [10.88, 12.36], $p = 0.002$), confirming the improvement is not isolated to a single Romance source language.

Model	chrF++ (fr → oc)	chrF++ (ca → oc)	cBPC (fr → oc)	BPC [HPLT]
`Simple Baseline`	30.88	29.76	**0.6403**	**1.1500**
`Full Pipeline`	**45.45**	**41.34**	0.6836	1.2712
Gap	+14.57	+11.58	+0.0433	+0.1212

Table 4.1: Translation quality (chrF++) and probabilistic metrics. Higher is better for chrF++, lower is better for BPC.

Moreover, the probabilistic metrics (cBPC and Unconditional BPC) show a regression for the full pipeline. The unconditional BPC evaluation across the Occitan corpora reveals that the performance gap is largest on the noisy web text (`hplt`: +0.1212), smaller on clean edited text (`ud`: +0.0561), and smallest on neutral prose (`flores`: +0.0449). This pattern mirrors the generalisation-versus-diversity tradeoff of the *alignment tax* (Ouyang et al., 2022; Kirk et al., 2023), where instruction tuning narrows output distributions around task-appropriate responses, inherently raising the Negative Log-Likelihood (NLL) on broader text distributions.

Additionally, as noted by Lin et al. (2023), alignment-induced distribution shifts frequently manifest around stylistic tokens. Because the full pipeline was strictly anchored to the Languedocien dialect via system prompting, its stylistic choices (e.g., specific clitics, elisions) diverge from the standard used in the FLORES-200 references. While chrF++ accommodates these dialectal paraphrases, exact-match sequence probability penalises them, explaining the specific cBPC regression.

4.2 Multilingual Ability via Romance-MoLE

This second evaluation showcases the Romance-MoLE architecture across a full 6-direction translation matrix to test the recovery of multilingual capabilities without catastrophic forgetting of Occitan. MoLE is compared against the Occitan-only upper bound (`FullPipeline`) and lower bound (`Simple`).

Target Direction	MoLE Final	Full Pipeline (Ceiling)	Simple (Floor)
oc ← fr	44.70	45.45	30.88
oc ← ca	43.75	41.34	29.76
fr ← oc	58.06	41.43	19.63
fr ← ca	55.02	37.63	19.74
ca ← oc	46.87	38.27	25.24
ca ← fr	34.02	37.28	26.47

Table 4.2: chrF++ results indicating target-language performance across models

As shown in Table 4.2, MoLE preserves Occitan performance near the expert ceiling, scoring an average of 44.23 across the two Occitan target directions (44.70 and 43.75) compared to the full expert's 43.40 (45.45 and 41.34). On the `ca` → `oc` task, MoLE surpassed the full pipeline (+2.42 chrF++, 95% CI [1.86, 2.94], $p = 0.002$), which

suggests that dynamic exposure to the Catalan expert assists the network in processing Catalan source inputs for Occitan translation.

The French targets demonstrate the most substantial multilingual gains. MoLE outperforms the single-language experts, improving by +16.63 chrF++ on `oc` → `fr` and +17.39 on `ca` → `fr` (both $p = 0.002$). Tokeniser-invariant metrics validate this improvement: MoLE achieved the best French cBPC (0.3729 and 0.3974) and the lowest unconditional BPC on `fr.flores` (0.8428), confirming the French generation ability.

To understand how the model routes these representations, the average gate mass per direction was extracted (Table 4.3).

Direction	Top Expert	Average Gate Mass
oc ← fr	`oc`	0.570
oc ← ca	`oc`	0.701
fr ← oc	`fr`	0.652
fr ← ca	`fr`	0.697
ca ← oc	`oc`	0.536
ca ← fr	`fr`	0.606

Table 4.3: MoLE routing attribution indicating the dominant expert and its gate mass per translation direction.

Routing attribution aligns with the quality of outputs for Occitan and French: generation heavily favours the respective target-language expert. However, the table highlights why Catalan translations yield mixed results. Neither Catalan target direction is dominated by a Catalan expert; `ca` ← `oc` routes mostly through Occitan, and `ca` ← `fr` routes mostly through French. Thus, while MoLE adds real Catalan capability compared to the baseline, its Catalan representation remains less cleanly isolated within the router.

4.3 Fine-Grained Grammatical Adherence

The third evaluation uses a synthetic minimal-pair challenge set to evaluate local morphosyntactic preferences. It specifically tests Languedocien Occitan forms (e.g., elisions, object clitics, and dialect shibboleths) against near-identical incorrect alternatives, scoring candidates via conditional Bits-Per-Character (cBPC). All models performed above the 50% random chance ($p < 0.0001$), suggesting the baseline training acquired general Occitan knowledge. As detailed in Table 4.4, the single-language models showed minimal variance, with the `Full Pipeline` scoring exactly one item higher overall than the `Simple` baseline (59/74 vs 58/74). This indicates that Evaluation 1's large chrF++ gains (Section 4.1) are driven by broader improvements in generation fluency rather than just the isolated morphosyntactic phenomena tested here.

The Romance-MoLE model (evaluated at a post-hoc sharpened router temperature of $\tau = 0.25$) achieved the highest overall accuracy of 81.1% (60/74 correct pairs, 95%

Grammatical Category	MoLE Final ($\tau = 0.25$)	Full Pipeline	Simple Baseline
Agreement	**10/10**	**10/10**	**10/10**
Articles	**10/10**	**10/10**	9/10
Auxiliaries (Tense)	1/4	1/4	1/4
Contractions	8/10	**9/10**	8/10
Elision	**9/10**	**9/10**	**9/10**
Languedocien Shibboleths	**10/10**	**10/10**	**10/10**
Negation	**4/10**	3/10	3/10
Object Clitics	**8/10**	7/10	**8/10**
Overall Accuracy	**81.1%** (60/74)	79.7% (59/74)	78.4% (58/74)

Table 4.4: Minimal-pair preference accuracy by grammatical category. The MoLE architecture preserves or slightly exceeds the fine-grained Occitan capabilities of the single-language experts.

CI [0.707, 0.884]). MoLE matched the single-expert models on specific Languedocien traits, scoring a perfect 10/10 on agreement, articles, and dialect shibboleths, and 9/10 on pre-vowel elisions. It marginally improved upon the full pipeline in handling negation (+1) and object clitics (+1). This performance suggests that multilingual routing preserves the Languedocien orthographic constraints acquired during the synthetic data curriculum.

On the other hand, the minimal-pair diagnostic also highlighted persistent base-model weaknesses across all architectures. As visible in Table 4.4, all models performed poorly on tense auxiliaries (1/4 correct) and only weakly on negation syntax ($\leq$ 4/10 correct). This indicates that while MoLE safeguards the grammar learned by the Occitan expert, certain complex temporal and negative features remain underrepresented in the training data and will require targeted synthetic curation in future iterations.

4.4 Routing Dynamics and HMoRA Analysis

To understand the mechanics of the Hierarchical Model Routing (HMoRA) strategy outlined in Section 3.8, expert activation heatmaps and qualitative generation outputs were analysed across different inferences. The analysis focused on how routing temperatures affect the generation quality and structural integrity of the output text.

Token-level and sequence-level routing exhibited different behaviours. Under default *soft routing*, token-level gates (Figure 4.1) displayed granular local mixing. While this theoretically allows the model to shift between the Occitan expert (for structural syntax) and the French or Catalan experts (for shared phonological subwords), qualitative tests revealed a major drawback: soft token-level routing frequently led to target-language drift and morphological collapse. Because each subword position receives an independent gate vector, the router can switch dominant expert mid-word, for instance, routing the stem of an Occitan verb through the French expert while routing the inflectional suffix through the Catalan expert, producing morphologically incoherent hybrid forms.

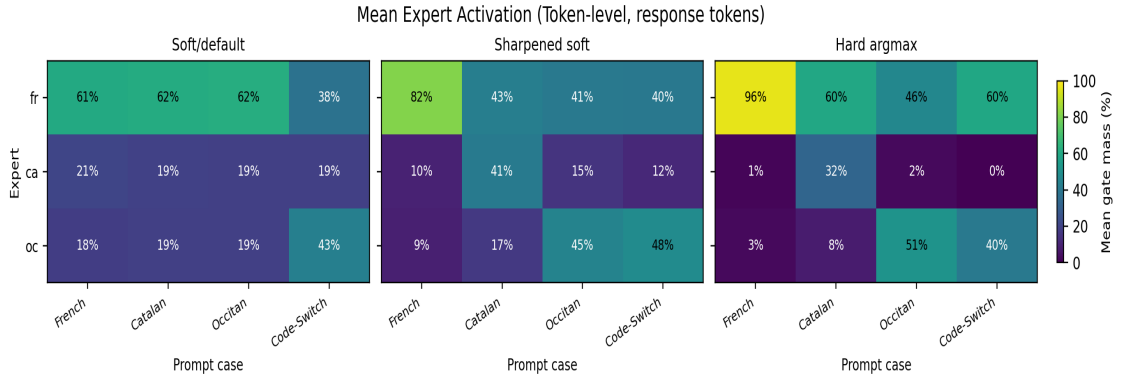

Figure 4.1: Mean Expert Activation (Token-level). The router exhibits granular mixing to process localised subword boundaries, which occasionally causes target-language drift under soft routing.

For example, when queried about a daily morning routine, Occitan and Catalan prompts suffered by having features from the other languages being blended together incoherently. Because French token boundaries dominated the internal states, the router underspecialised, ultimately producing fluent but entirely incorrect French outputs for non-French prompts (e.g., *"Je me réveille à 6h30 du matin..."*).

Applying *sharpened soft routing*, natively supported by HMoRA's sequence-level mean-pooling in the deeper transformer layers, mitigated this erroneous code-switching. By pooling gate probabilities in the deep layers, the network is forced into a coherent, document-level language decision (Figure 4.2).

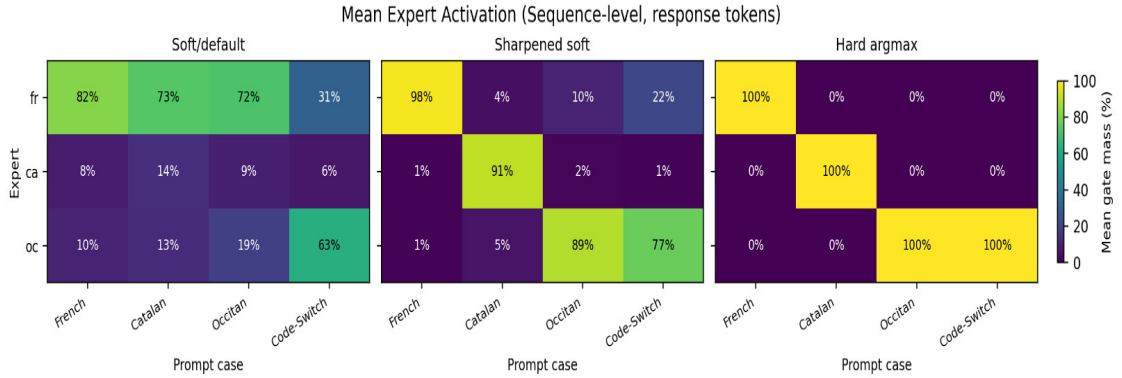

Figure 4.2: Mean Expert Activation (Sequence-level). Sharpened soft routing focuses sequence-level mass on the target language, effectively mitigating mid-sentence code-switching.

The impact of this hierarchical sharpening is clear. When the router is adjusted to enforce sequence-level targeting, the model recovers target-language pragmatics and

fine-grained morphology. Table 4.5 illustrates the generated outputs under sharpened conditions.

Input Target	**Top Expert Gate**	**Generated Output Excerpt (Sharpened Routing)**
French	`fr` (~ 0.98)	*"Je me réveille à 6 heures du matin, je me lève..."*
Catalan	`ca` / `fr`	*"Com a assistent virtual, la meva rutina del matí..."*
Occitan	`oc` (~ 0.89)	*"Abans d'entrar al trabalh, soi l'assistant virtual..."*
Code-Switch	`oc`	*"La ville de Sant Gaudenç es situada dins la vallada..."*

Table 4.5: Qualitative generation outputs across different input targets using sharpened sequence-level routing.

In the Occitan output, the sequence-level router assigns approximately 89% of its probability mass to the Occitan expert. Consequently, the output demonstrates high-quality Languedocien syntax (*"Abans d'entrar al trabalh, soi l'assistant virtual..."*). The linguistic quality here is noticeable, the model correctly applies mandatory pre-vowel elisions (*d'entrar*, *l'assistant*), accurate first-person verbal conjugations (*soi*), and appropriate regional vocabulary (*trabalh*). It avoids the French calques and structural bleeding that affected the default soft-routing configurations. Similarly, the Catalan prompt maps to native grammatical structures (*"Com a assistent virtual, la meva rutina..."*).

The architecture proves robust to noisy, code-switched inputs. When provided with a mixed Romance prompt designed to confuse the model, it routes the majority of its mass through the Occitan expert to stabilise the generation (*"La ville de Sant Gaudenç es situada dins la vallada..."*). Despite the presence of high-resource French distractors in the input (as seen in the spelling of *ville*), the HMoRA structure aligns the underlying grammatical framework to Occitan (*es situada dins la*). Thus, the output confirms that Romance-MoLE facilitates multilingual switching without compromising the structural and grammatical integrity of the generated output.

5 Conclusion

5.1 Discussion

The primary objective of this thesis was to establish a parameter-efficient, scalable pipeline for integrating low-resource languages into Large Language Models. Specifically, we focus on the Languedocien dialect of Occitan, achieving this integration, while reducing the effects of the "Curse of Multilinguality" (Conneau et al., 2020). The empirical results demonstrate that this is not only possible but effective when using targeted cross-lingual transfer from phylogenetic close, high-resource languages.

The +14.57 chrF++ improvement of the `Full Pipeline` over the simple baseline in Evaluation 1 (Section 4.1) validates this multi-stage approach. Naive fine-tuning on web-scraped fragmented corpora is insufficient for marginalised languages. However, by repairing the tokeniser boundaries via Trans-Tokenisation and initialising Occitan through the TIES-Merging of French and Catalan LoRA experts, we mapped low-resource morphology onto the model's pre-existing high-resource logic. The observed "alignment tax" where the model traded some unconditional probabilistic accuracy (BPC) in favour of strict, system-prompted dialect alignment proves to be a necessary trade-off. It indicates that the model learned to reject high-resource structural calques in favour of Languedocien forms.

Additionally, the Romance-MoLE architecture minimised the catastrophic parameter interference that typically affects multilingual models. By enforcing Hierarchical Model Routing (HMoRA), the model mitigated the mid-sentence code-switching frequently observed in shallow, token-level routing regimes. As shown in (Section 4.2), Romance-MoLE not only maintained Occitan performance near the isolated expert ceiling but also recovered massive generative capabilities in French (+16.63 chrF++).

These findings prove that digital isolation is not an inevitable fate for marginalised languages. By framing language preservation as a combined architectural routing and data scaling problem, we have established a framework. Minority languages do not need to be mathematically assimilated in shared latent spaces. Instead, with careful phylogenetic scaffolding, they can coexist alongside dominant global languages.

5.2 Limitations

Despite the improvement demonstrated by the Romance-MoLE framework, several structural and methodological limitations remain.

Synthetic Data and High-Resource Bias To overcome the data scarcity of the HPLT corpus, this research relied on synthetic data generated by high-resource models (e.g., the Gemini API). While system prompts enforced the Languedocien *Norma Classica*, qualitative expert evaluations highlighted that high-resource biases inevitably seeped into the model's stylistic phrasing. The occasional reliance on French compound past tenses and Catalan lexical mappings indicates that "translationese" persists when utilising cross-lingual transfer from dominant languages. Furthermore, lexical hallucinations in out-of-distribution domains (e.g., specific nature vocabulary) expose the limitations of relying on synthetic generation for culturally localised terminology. Crucially, these challenges are exacerbated by a lack of reliable access to native speakers. Without native Languedocien annotators to verify the quality of the synthetic data and evaluate the models' outputs, capturing and iteratively correcting these subtle linguistic biases remains an ongoing challenge.

Grammatical Blindspots While the Romance-MoLE architecture excelled at morphological adherence (e.g., elisions and dialectal shibboleths scoring 10/10), Evaluation 3 (Section 4.3) exposed structural weaknesses in deeper syntax. All tested models performed poorly on tense auxiliaries and complex negation. This suggests that while surface-level alignment can enforce orthography and clitic boundaries, modelling complex temporal and negative features requires a denser, more structurally diverse training curriculum than our synthetic data generation budget allowed.

Routing Boundaries for Proximate Languages The MoLE routing attribution revealed that while Occitan and French representations were distinctly isolated, Catalan routing remained fuzzy. Catalan targets were frequently routed through Occitan or French experts rather than the dedicated Catalan adapter. This indicates that because Catalan and Occitan share close phylogenetic proximity, and because the base model (`Llama-3.1-Carballo`) possesses a massive underlying Ibero-Romance bias, the router struggled to penalise and decouple their internal representations.

Compute-Constrained Architecture The methodology was bound by a strict 48GB VRAM hardware limit. Consequently, the pipeline relied entirely on Low-Rank Adaptation. While efficient, freezing the majority of the base model's parameters restricts the network's total representational capacity compared to full-parameter continued pretraining.

5.3 Future Work

The framework established in this thesis opens several avenues for future research, both in computational architecture and broader linguistic preservation efforts.

Expanding the Synthetic Curriculum Future iterations must address the grammatical blindspots identified during the minimal-pair evaluations. Expanding the synthetic generation pipeline to include complex structures, such as the Occitan preterit, subordinate clauses, and varied negation patterns, researchers can create a more robust training foundation. Additionally, expanding the domain vocabulary beyond conversational ALPACA prompts into STEM, historical, and ecological texts will likely help with the lexical hallucination rates observed.

Refining MoLE Router Isolation To resolve the ambiguous routing boundaries between tightly clustered languages like Occitan and Catalan, future work should explore contrastive routing penalties or dynamic rank allocation. Training the router with explicit negative samples (e.g., penalising Catalan expert activation during Occitan generation and vice versa) could decouple these closely-related representations without losing the benefits of their shared syntactic roots.

Mapping the Broader Dialect Continuum This thesis specifically targeted the Languedocien dialect to establish a standardised computational baseline. A natural next step is to expand the Romance-MoLE architecture to dynamically route between the broader Occitan dialect continuum. By training individual, low-rank experts for individual Occitan dialects (e.g., Gascon, Provençal, and Limousin), the MoLE framework could theoretically allow a single LLM to navigate internal regional variations without homogenising them into a single, artificial standard (Bird, 2020).

Applying the Framework to Other Endangered Languages Lastly, the ultimate goal of this pipeline is community adaptability. Future research should apply the TransTokenisation and TIES-Merging methodologies to other marginalised languages that share similar phylogenetic environments. Expanding this framework to languages like Aromanian, Sardinian, or indigenous languages with high-resource linguistic neighbours will test the true generalisability of the "Rich Cousins" cross-lingual transfer strategy.

5.4 Concluding Remarks

As LLMs rapidly become the fundamental infrastructure of the modern digital world, ensuring the inclusion of under-resourced languages is an ethical imperative. The sociopolitical marginalisation of languages like Occitan and my own native Aromanian must not be compounded by digital erasure. By creatively combining state-of-the-art parameter-efficient fine-tuning, dynamic routing, and phylogenetic scaffolding, this thesis demonstrates a viable path forward. It transforms LLMs from agents of linguistic

assimilation into accessible tools for cultural preservation, giving marginalised communities a chance to survive and thrive in the modern digital era.

Bibliography

Aepli, Noëmi, Chantal Amrhein, Florian Schottmann, and Rico Sennrich (Dec. 2023). "A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz. Singapore: Association for Computational Linguistics, pp. 1045–1065. DOI: 10.18653/v1/2023.wmt-1.99. URL: https://aclanthology.org/2023.wmt-1.99/.

Ahia, Orevaoghene, Hao Jiang, Kelechi Ogueji, and Sara Hooker (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*.

AIffl (2024). *Alpaca_french_mixtral Dataset*. URL: https://huggingface.co/datasets/AIffl/Alpaca_french_mixtral.

Alibert, Loís (1935). *Gramatica occitana segon los parlars lengadocians*. Tolosa: Societat d'Estudis Occitans.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (Nov. 2020). "Translation Artifacts in Cross-lingual Transfer Learning". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. URL: https://aclanthology.org/2020.emnlp-main.618/.

Bird, Steven (2020). "Decolonising Speech and Language Technology". English. In: *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference. Association for Computational Linguistics (ACL), pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313.

Blanchet, Philippe (2004). "Provençal as a distinct language? Sociolinguistic patterns revealed by a recent public and political debate". In.

— (2018). "Discriminations: combattre la glottophobie". In: *Plurilinguisme, entreprises, économie et société*. Observatoire européen du plurilinguisme, pp. 209–219.

Boukhari, T. (2023). *Alpaca_french_instruct Dataset*. URL: https://huggingface.co/datasets/tboukhari/Alpaca_french_instruct.

Buehler, Eric L. and Markus J. Buehler (2024). *X-LoRA: Mixture of Low-Rank Adapter Experts, a Flexible Framework for Large Language Models with Applications in Protein Mechanics and Molecular Design*. arXiv: 2402.07148 [cond-mat.soft]. URL: https://arxiv.org/abs/2402.07148.

Burchell, Laurie et al. (2025). "An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT)". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. URL: https://aclanthology.org/2025.acl-long.854/.

Chang, Tyler A., Catherine Arnett, Zhuowen Tu, and Benjamin Bergen (2024). "When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages". In: *arXiv preprint arXiv:2311.09205*. URL: https://arxiv.org/abs/2311.09205.

Chen, Tianqi, Bing Xu, Chiyuan Zhang, and Carlos Guestrin (2016). "Training Deep Nets with Sublinear Memory Cost". In: *arXiv preprint arXiv:1604.06174*. DOI: 10.48550/arxiv.1604.06174.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747/.

Costa, James (2023). "A materialist take on minoritization, emancipation, and language revitalization: Occitan sociolinguistics since the 1970s". In: *Journal of sociolinguistics* 27.4, pp. 327–344.

Costa-jussà, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation". In: *arXiv preprint arXiv:2207.04672*.

Dao, Tri (2023). "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning". In: *arXiv preprint arXiv:2307.08691*. DOI: 10.48550/arxiv.2307.08691.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *Advances in Neural Information Processing Systems*. Vol. 36. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb8787143051f92683944a99774ba3-Abstract-Conference.html.

Etxaniz, Julen, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe (June 2024). "Do Multilingual Language Models Think Better in English?" In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 550–564. DOI: 10.18653/v1/2024.naacl-short.46. URL: https://aclanthology.org/2024.naacl-short.46/.

Fedus, William, Barret Zoph, and Noam Shazeer (2022). "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity". In: *Journal of Machine Learning Research* 23.120, pp. 1–39. URL: http://jmlr.org/papers/v23/21-0998.html.

Ferguson, Charles A. (1959). "Diglossia". In: *WORD* 15.2, pp. 325–340. DOI: 10.1080/00437956.1959.11659702. eprint: https://doi.org/10.1080/00437956.1959.11659702. URL: https://doi.org/10.1080/00437956.1959.11659702.

Fishman, Joshua A. (1967). "Bilingualism With and Without Diglossia; Diglossia With and Without Bilingualism". In: *Journal of Social Issues* 23.2, pp. 29–38. DOI: https://doi.org/10.1111/j.1540-4560.1967.tb00573.x. eprint: https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4560.1967.tb00573.x. URL: https://spssi.onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4560.1967.tb00573.x.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky (2016). "Domain-Adversarial Training of Neural Networks". In: *Journal of Machine Learning Research* 17.59, pp. 1–35. URL: http://jmlr.org/papers/v17/15-239.html.

Google DeepMind (2025a). *Gemini 3 Flash Model Card*. Tech. rep. Google DeepMind. URL: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

— (2025b). *Gemini 3 Pro Model Card*. Tech. rep. Google DeepMind. URL: https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Grattafiori, Aaron et al. (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

Gudibande, Arnav, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song (2023). *The False Promise of Imitating Proprietary LLMs*. arXiv: 2305.15717 [cs.CL]. URL: https://arxiv.org/abs/2305.15717.

Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. (2023). "Textbooks are all you need". In: *arXiv preprint arXiv:2306.11644*.

Harris, Martin and Nigel Vincent (2003). *The romance languages*. Routledge.

Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, et al. (2022). "Challenges and Strategies in Cross-Cultural NLP". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. URL: https://aclanthology.org/2022.acl-long.482.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015). *Distilling the Knowledge in a Neural Network*. arXiv: 1503.02531 [stat.ML]. URL: https://arxiv.org/abs/1503.02531.

Hopton, Zachary and Noëmi Aepli (June 2024). "Modeling Orthographic Variation in Occitan's Dialects". In: *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*. Ed. by Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov, and Jörg Tiedemann. Mexico City, Mexico: Association for Computational Linguistics, pp. 78–88. DOI: 10.18653/v1/2024.vardial-1.6. URL: https://aclanthology.org/2024.vardial-1.6/.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (2019). *Parameter-Efficient Transfer Learning for NLP*. arXiv: 1902.00751 [cs.LG]. URL: https://arxiv.org/abs/1902.00751.

Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2022). "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

Ilharco, Gabriel, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi (2023). *Editing Models with Task Arithmetic*. arXiv: 2212.04089 [cs.LG]. URL: https://arxiv.org/abs/2212.04089.

Inan, Hakan, Khashayar Khosravi, and Richard Socher (2016). "Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling". In: *arXiv preprint arXiv:1611.01462*.

Jain, Neel, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Yeh, Avi Schwarzschild, and Tom Goldstein (2023). "NEFTune: Noisy Embeddings Improve Instruction Finetuning". In: *arXiv preprint arXiv:2310.05914*.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (July 2020). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 6282–6293. DOI:

10.18653/v1/2020.acl-main.560. URL: https://aclanthology.org/2020.acl-main.560.

Kirk, Robert, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu (2023). "Understanding the effects of rlhf on llm generalisation and diversity". In: *arXiv preprint arXiv:2310.06452*.

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell (Mar. 2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13, 3521–3526. ISSN: 1091-6490. DOI: 10.1073/pnas.1611835114. URL: http://dx.doi.org/10.1073/pnas.1611835114.

Koehn, Philipp (Jan. 2004). "Statistical Significance Tests for Machine Translation Evaluation." In: pp. 388–395.

Komatsuzaki, Aran, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby (2023). *Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints*. arXiv: 2212.05055 [cs.LG]. URL: https://arxiv.org/abs/2212.05055.

Koppel, Moshe and Noam Ordan (June 2011). "Translationese and Its Dialects". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, pp. 1318–1326. URL: https://aclanthology.org/P11-1132/.

Kunz, Jenny, Iben Nyholm Debess, and Annika Simonsen (2025). "Family Matters: Language Transfer and Merging for Adapting Small LLMs to Faroese". In: *arXiv preprint arXiv:2510.00810*. URL: https://arxiv.org/abs/2510.00810.

Kuparinen, Olli, Aleksandra Miletić, and Yves Scherrer (Dec. 2023). "Dialect-to-Standard Normalization: A Large-Scale Multilingual Evaluation". In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 13814–13828. DOI: 10.18653/v1/2023.findings-emnlp.923. URL: https://aclanthology.org/2023.findings-emnlp.923/.

Köpf, Andreas, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick (2023). *OpenAssistant Conversations – Democratizing Large Language Model Alignment*. arXiv: 2304.07327 [cs.CL]. URL: https://arxiv.org/abs/2304.07327.

Lafont, Robert (1997). "Quarante ans de sociolinguistique à la périphérie". In.

Li, Zhuoyan, Hangxiao Zhu, Zhuoran Lu, and Ming Yin (2023). *Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations*. arXiv: 2310.07849 [cs.CL]. URL: https://arxiv.org/abs/2310.07849.

Liao, Mengqi, Wei Chen, Junfeng Shen, Shengnan Guo, and Huaiyu Wan (2025). "HMoRA: Making LLMs More Effective with Hierarchical Mixture of LoRA Experts". In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=lTkHiXeuDl.

Lin, Bill Yuchen, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi (2023). "The unlocking spell on base llms: Rethinking alignment via in-context learning". In: *arXiv preprint arXiv:2312.01552*.

Longpre, Shayne, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, and Sayna Ebrahimi (2026). *ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality*. arXiv: 2510.22037 [cs.CL]. URL: https://arxiv.org/abs/2510.22037.

McCloskey, Michael and Neal J. Cohen (1989). "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". In: ed. by Gordon H. Bower. Vol. 24. Psychology of Learning and Motivation. Academic Press, pp. 109–165. DOI: https://doi.org/10.1016/S0079-7421(08)60536-8. URL: https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Moseley, Christopher and Alexandre Nicolas (2010). *UNESCO Interactive Atlas of the World's Languages in Danger*. Archived from the original on June 27, 2012. URL: https://web.archive.org/web/20120627041634/http://www.unesco.org/culture/languages-atlas/index.php.

Nédey, Oriane, Rachel Bawden, Thibault Clérice, and Benoît Sagot (Mar. 2026). "OcWikiDialects: A Wikipedia Dataset With Rich Metadata for Occitan Dialect Identification". In: *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*. Ed. by Yves Scherrer, Noëmi Aepli, Verena Blaschke, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. Rabat, Morocco: Association for Computational Linguistics, pp. 45–57. DOI: 10.18653/v1/2026.vardial-1.4. URL: https://aclanthology.org/2026.vardial-1.4/.

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi

Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir (Nov. 2020). "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195. URL: https://aclanthology.org/2020.findings-emnlp.195/.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744.

Pacifico, Jonathan (2024). *French-Alpaca-dataset-Instruct-110K*. https://huggingface.co/datasets/jpacifico/French-Alpaca-dataset-Instruct-110K.

Pearce, Tim (2023). *alpaca-cleaned-french*. https://huggingface.co/datasets/timpearce/alpaca-cleaned-french.

Pecher, Branislav, Jan Cegin, Robert Belanec, Ivan Srba, Jakub Simko, and Maria Bielikova (2026). "Better as Generators Than Classifiers: Leveraging LLMs and Synthetic Data for Low-Resource Multilingual Classification". In: *arXiv preprint arXiv:2601.16278*.

Petrov, Aleksandar, Emanuele La Malfa, Philip Torr, and Adel Bibi (2023). "Language Model Tokenizers Introduce Unfairness Between Languages". In: *arXiv preprint arXiv:2305.15425*.

Pfeiffer, Jonas, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych (Apr. 2021). "AdapterFusion: Non-Destructive Task Composition for Transfer Learning". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 487–503. DOI: 10.18653/v1/2021.eacl-main.39. URL: https://aclanthology.org/2021.eacl-main.39/.

Press, Ofir and Lior Wolf (Apr. 2017). "Using the Output Embedding to Improve Language Models". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 157–163. URL: https://aclanthology.org/E17-2025/.

Pym, Anthony (2004). *The Moving Text: Localization, Translation, and Distribution*. John Benjamins Publishing.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He (2020). *ZeRO: Memory Optimizations Toward Training Trillion Parameter Models*. arXiv: 1910.02054 [cs.LG]. URL: https://arxiv.org/abs/1910.02054.

Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur (Feb. 2023). "Neural Machine Translation for Low-resource Languages: A Survey". In: *ACM Comput. Surv.* 55.11. ISSN: 0360-0300. DOI: 10.1145/3567592. URL: https://doi.org/10.1145/3567592.

Remy, François, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester (2024). *Trans-Tokenization and Cross-Lingual Vocabulary Transfers: Language Adaptation of LLMs for Low-Resource NLP*. arXiv: 2408.04303 [cs.CL]. URL: https://arxiv.org/abs/2408.04303.

Robinson, Nathaniel, Perez Ogayo, David R. Mortensen, and Graham Neubig (Dec. 2023). "ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages". In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz. Singapore: Association for Computational Linguistics, pp. 392–418. DOI: 10.18653/v1/2023.wmt-1.40. URL: https://aclanthology.org/2023.wmt-1.40/.

Rodríguez, Pablo, Silvia Paniagua Suárez, Pablo Gamallo, and Susana Sotelo Docio (2025). "Continued Pretraining and Interpretability-Based Evaluation for Low-Resource Languages: A Galician Case Study". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics. URL: https://aclanthology.org/2025.findings-acl.240/.

Sakajo, Haruki, Yusuke Ide, Justin Vasselli, Yusuke Sakai, Yingtao Tian, Hidetaka Kamigaito, and Taro Watanabe (2025). "Dictionaries to the Rescue: Cross-Lingual Vocabulary Transfer for Low-Resource Languages Using Bilingual Dictionaries". In: *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25963–25976. URL: https://aclanthology.org/2025.findings-acl.1333.

Schut, Lisa, Yarin Gal, and Sebastian Farquhar (2025). *Do Multilingual LLMs Think In English?* arXiv: 2502.15603 [cs.CL]. URL: https://arxiv.org/abs/2502.15603.

Schöffel, Matthias, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and Matthias Aßenmacher (2025). *Modern Models, Medieval Texts: A POS Tagging Study of Old Occitan*. arXiv: 2503.07827 [cs.CL]. URL: https://arxiv.org/abs/2503.07827.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto (2023). *Stanford Alpaca: An Instruction-following LLaMA model*. URL: https://github.com/tatsu-lab/stanford_alpaca.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample (2023).

LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

Upadhayay, Bibek and Vahid Behzadan (2024). "TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes". In: *5th Workshop on practical ML for limited/low resource settings, ICLR*. URL: https://openreview.net/forum?id=02MLWBj8HP.

van Vugt, Jan (2018). *pytorch-domain-adaptation*. URL: https://github.com/jvanvugt/pytorch-domain-adaptation.

Wang, Kuan, Alexander Bukharin, Xiaojian Ma, Zeyu Liu, Ruiqi Zhong, Denny Zhou, Shihao Ji, Mengdi Wang, and Shiyang Li (2024a). *RNR: Teaching Large Language Models to Follow Roles and Rules*. arXiv: 2409.13733 [cs.CL]. URL: https://arxiv.org/abs/2409.13733.

Wang, Lean, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai (2024b). *Auxiliary-Loss-Free Load Balancing Strategy for Mixture-of-Experts*. arXiv: 2408.15664 [cs.LG]. URL: https://arxiv.org/abs/2408.15664.

Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi (2023). *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. arXiv: 2212.10560 [cs.CL]. URL: https://arxiv.org/abs/2212.10560.

Wang, Zirui, Zachary C. Lipton, and Yulia Tsvetkov (Nov. 2020). "On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 4438–4450. DOI: 10.18653/v1/2020.emnlp-main.359. URL: https://aclanthology.org/2020.emnlp-main.359/.

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West (Aug. 2024). "Do Llamas Work in English? On the Latent Language of Multilingual Transformers". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 15366–15394. DOI: 10.18653/v1/2024.acl-long.820. URL: https://aclanthology.org/2024.acl-long.820/.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, et al. (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45.

Woller, Lisa, Viktor Hangya, and Alexander Fraser (2021). "Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings". In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 41–50.

Xu, Benfeng, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yin Zhang, and Zhendong Mao (2023). *ExpertPrompting: Instructing Large Language Models to be Distinguished Experts*. arXiv: 2305.14688 [cs.CL]. URL: https://arxiv.org/abs/2305.14688.

Yadav, Prateek, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal (2023). "TIES-Merging: Resolving Interference When Merging Models". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=xtaX3WyCj1.

Zadouri, Ted, Ahmet Unal, Vinitra Swamy, Jibril Frej, Mary-Anne Hartley, Antoine Bosselut, and Martin Jaggi (2024). "Pushing Mixture of Experts to the Limit: Extremely Parameter Efficient MoE for Instruction Tuning". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=6d00071564.

Zhao, Yanli, Andrew Gu, Rohan Vujanic, et al. (2023). "PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel". In: *Proceedings of the VLDB Endowment* 16.12, pp. 3848–3860.

Zhou, Chunting, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy (2023). "LIMA: Less Is More for Alignment". In: *Advances in Neural Information Processing Systems*. URL: https://arxiv.org/abs/2305.11206.

Appendices

A Prompts and Dataset Sample

This appendix records the alignment-steering logic and small curated data excerpts used in the Occitan adaptation pipeline.

A.1 Gemini Synthetic Generation Prompts

A.1.1 FLORES-200 French to Languedocien Occitan

Model: `gemini-3-pro-preview`
The generation script used the following system instruction verbatim:

```
You are an expert translator and linguist specializing in Occitan,
specifically the Languedocien (Lengadocian) dialect following the Classical
Norm (Norma Classica) established by Louis Alibert.

Your task is to **localize** French text into Languedocien Occitan. This is
NOT a literal word-for-word translation|you must produce natural, idiomatic
Occitan that a native speaker of the Lengadocian dialect would use.

## STRICT GRAMMATICAL RULES (Norma Classica - Alibert):

### 1. Definite Articles:
- USE: lo (masc. sing.), la (fem. sing.), los (masc. plur.), las (fem. plur.)
- DO NOT USE Gascon articles: eth, era, eths, eras

### 2. Contractions and Prepositions:
- USE: del (de + lo), pel (per + lo), al (a + lo)
- USE: dels, pels, als for plurals
- DO NOT USE Provençal contractions: dau, daus, pòu
```

```
### 3. Elision (MANDATORY - Apply Strictly):
- Apply elision before vowels and silent 'h':
  - lo/la → l' (l'òme, l'aiga)
  - de → d' (d'aquò, d'aquel)
  - que → qu' (qu'es, qu'avèm)
  - se → s' (s'escapa)
  - ne → n' (n'i a)
  - me, te → m', t' (m'agrada, t'ai vist)

### 4. Apostrophe Style:
- USE ONLY straight apostrophes: '
- DO NOT USE curly/typographic apostrophes: ‘ or ’

### 5. Spelling Conventions:
- Use 'ò' for open 'o' sounds
- Use 'è' for open 'e' sounds
- Use 'ç' before a, o, u for /s/ sound
- Use 'nh' for palatal nasal (like Spanish 'ñ')
- Use 'lh' for palatal lateral (like Italian 'gl' in famiglia)

### 6. Verb Forms (Languedocien conjugation):
- Follow Alibert's verb conjugation tables
- Use -èm for 1st person plural present (parlam → parlam is also acceptable)
- Use -ètz for 2nd person plural
- Use -on/-an for 3rd person plural depending on conjugation class

## OUTPUT REQUIREMENTS:
- Provide ONLY the Occitan translation, nothing else
- No explanations, notes, or alternatives
- No quotation marks around the translation
- Preserve the original meaning and tone
- Adapt cultural references appropriately when needed
```

The per-example user prompt was:

```
Translate the following French text into Languedocien Occitan:

{french_text}
```

A.1.2 Catalan Alpaca to Languedocien Occitan

Model: `gemini-3-flash-preview` (with optional thinking configuration in the script). The generation script used the following system instruction verbatim:

```
You are an expert translator and linguist specializing in Occitan,
specifically the Languedocien (Lengadocian) dialect following the Classical
Norm (Norma Classica) established by Louis Alibert.

Your task is to **localize** Catalan text into Languedocien Occitan. This is
NOT a literal word-for-word translation|you must produce natural, idiomatic
Occitan that a native speaker of the Lengadocian dialect would use. Catalan
and Occitan are closely related; preserve the meaning and adapt spelling and
grammar to Norma Classica.

## STRICT GRAMMATICAL RULES (Norma Classica - Alibert):

### 1. Definite Articles:
- USE: lo (masc. sing.), la (fem. sing.), los (masc. plur.), las (fem. plur.)
- DO NOT USE Gascon articles: eth, era, eths, eras

### 2. Contractions and Prepositions:
- USE: del (de + lo), pel (per + lo), al (a + lo)
- USE: dels, pels, als for plurals
- DO NOT USE Provençal contractions: dau, daus, pòu

### 3. Elision (MANDATORY - Apply Strictly):
- Apply elision before vowels and silent 'h':
  - lo/la → l' (l'òme, l'aiga)
  - de → d' (d'aquò, d'aquel)
  - que → qu' (qu'es, qu'avèm)
  - se → s' (s'escapa)
  - ne → n' (n'i a)
  - me, te → m', t' (m'agrada, t'ai vist)

### 4. Apostrophe Style:
- USE ONLY straight apostrophes: '
- DO NOT USE curly/typographic apostrophes: ‘ or ’

### 5. Spelling Conventions:
- Use 'ò' for open 'o' sounds
- Use 'è' for open 'e' sounds
- Use 'ç' before a, o, u for /s/ sound
- Use 'nh' for palatal nasal (like Spanish 'ñ')
- Use 'lh' for palatal lateral (like Italian 'gl' in famiglia)

### 6. Verb Forms (Languedocien conjugation):
- Follow Alibert's verb conjugation tables
```

```
- Use -èm for 1st person plural present
- Use -ètz for 2nd person plural
- Use -on/-an for 3rd person plural depending on conjugation class

## CONTENT ADAPTATION RULES (CRITICAL):

### 7. Cultural Localization:
- Replace references to non-Occitan geography, companies, or laws with
  Occitan/Southern-French equivalents when possible.
- If a concept is universally applicable (science, math, cooking), keep it
  but express it naturally in Occitan.
- Do NOT invent Occitan-specific facts. If the content cannot be meaningfully
  adapted, translate it faithfully without adding extra commentary.

### 8. Forbidden Output Patterns:
- NEVER include URLs, web links, or references to websites.
- NEVER include footnote-style references like [1], [2], etc.
- NEVER include "Works Cited", bibliographies, or source attributions.
- NEVER mention social media platforms (Facebook, Instagram, Twitter, etc.).
- NEVER add translations into other languages (Spanish, Italian, etc.).
- NEVER say "as an AI" or disclaim inability to access real-time data.
- NEVER reference the Merriam-Webster dictionary or any encyclopedia by name.
- Keep responses self-contained: the answer must stand on its own without
  pointing the reader elsewhere.

## OUTPUT REQUIREMENTS:
- You will receive a Catalan Alpaca example with three fields: instruction,
  input, output.
- Return a valid JSON object with exactly three keys: "instruction", "input",
  "output".
- Each value must be the Occitan localization of the corresponding Catalan text.
- If the original "input" is empty or "nan", return "" for "input".
- Preserve the original meaning and tone; adapt cultural references when
  appropriate.
- No explanations, no markdown code fences|only the raw JSON object.
```

The per-example user prompt was:

```
Localize this Catalan Alpaca example into Languedocien Occitan. Return only
a JSON object with keys "instruction", "input", "output".

Catalan:
- instruction: {instruction}
- input: {input_display}
```

```
- output: {output_text}

Occitan (JSON only):
```

A.1.3 Morphological Stress-Test Generation

Model: `gemini-3-pro-preview`
The stress-test generator used a compact global system instruction and category-specific prompts. The global instruction was:

```
You are a computational linguist generating training data for an Occitan
tokeniser.

Your role is to produce sentences that follow EXACT morphological specifications.
You must output ONLY the requested Occitan sentences in valid JSON format.

CRITICAL RULES:
1. Output ONLY valid JSON - no explanations, no translations, no notes
2. Each sentence must be in Languedocien Occitan (Classical Norm / Norma Classica)
3. Follow the specific morphological requirements given in each prompt EXACTLY
4. Use straight apostrophes (') only, never curly quotes
5. Sentences must be grammatically correct and semantically coherent
6. Each sentence must be UNIQUE - no duplicates

You are NOT translating. You are GENERATING authentic Occitan text for NLP
training.
```

The `elision_overload` category prompt was:

```
Generate {count} distinct sentences in Languedocien Occitan.

RULES (MANDATORY):
- Each sentence MUST contain AT LEAST 3-4 elisions
- Target elision patterns: d', l', qu', s', n', m', t'
- Sentences should be natural but elision-dense

EXAMPLES of elision-heavy sentences:
- "S'es n'anat d'aquí sens dire d'ont veniá."
- "L'òme qu'aviá vist s'es escapat d'aquela traïna."
- "N'i a pas qu'un qu'o sabiá, e s'es n'anat."
- "M'an dit qu'es l'ora d'anar s'adormir."
- "T'ai vist qu'anaves d'ont veniá l'aiga."

Return ONLY a JSON array of sentences, no explanations:
["sentence1", "sentence2", ...]
```

The diacritic_density category prompt was:

```
Generate {count} distinct sentences in Languedocien Occitan.

RULES (MANDATORY):
- Sentences MUST be packed with diacritic-heavy words
- Required characters: ò, à, è, é, í, ú, ç
- Each sentence should contain at least 4-5 accented words
- Focus on words with open vowels (ò, è) and cedilla (ç)

EXAMPLES of diacritic-dense sentences:
- "Çò que vòls es la vertat sus l'istòria d'aquela bòria."
- "L'òme vièlh parlèt de la tèrra e de l'òrt pròche."
- "Aquò's la bèla cançon qu'ausiguèrem ièr al sèr."
- "Lo còr de la vilòta èra plen de gènt que cantava."
- "La fèsta comencèt amb una dança e un còp de canòn."

Return ONLY a JSON array of sentences, no explanations:
["sentence1", "sentence2", ...]
```

The shibboleth category prompt was:

```
Generate {count} distinct sentences in Languedocien Occitan.

RULES (MANDATORY):
- Focus on Languedocien-specific features that distinguish it from Gascon
  and Catalan
- USE definite articles: lo, la, los, las (NEVER Gascon "eth/era" or Catalan "el")
- USE contractions: del, pel, al, dels, pels, als
- Include characteristic Languedocien vocabulary

EXAMPLES of dialect-marking sentences:
- "Lo lop e la lèbre son dins lo bòsc."
- "Los mainatges an passat pel pont e son anats al mercat."
- "La femna del vailet trabalhava dins los camps."
- "Las flors del prat son las mai bèlas de la valada."
- "Al ser, lo solelh se cocha darrièr los tucs."

NEVER USE:
- Gascon articles: eth, era, eths, eras
- Catalan articles: el, els
- Gascon enunciative "Que" at sentence start

Return ONLY a JSON array of sentences, no explanations:
["sentence1", "sentence2", ...]
```

A.2 Cognate Alignment Constraints

Model: `gemini-3-pro-preview`
The token-alignment stage used Gemini to map newly introduced Occitan tokens to close French and Catalan forms before weighted embedding initialization. The full system instruction was:

```
You are an expert Romance linguist specializing in Occitan, French, and Catalan.

Your task is to translate Occitan words into their closest French and Catalan
equivalents.

## CRITICAL RULES:

### 1. Morphological Preservation
When an Occitan word contains contractions, elisions, or specific morphological
markers:
- Find the STRUCTURALLY EQUIVALENT form in French/Catalan
- Examples:
  - d'aquesta (de + aquesta) → French: "de cette" OR "d'une" / Catalan: "d'aquesta"
  - l'òme (l' + òme) → French: "l'homme" / Catalan: "l'home"
  - qu'es (qu' + es) → French: "qu'est" OR "qui est" / Catalan: "que és"

### 2. Single Words
For simple words without contractions:
- Provide the most common translation
- Examples:
  - òme → French: "homme" / Catalan: "home"
  - femna → French: "femme" / Catalan: "dona"

### 3. Function Words
For articles, prepositions, conjunctions:
- Provide the direct equivalent
- Examples:
  - lo → French: "le" / Catalan: "el"
  - del → French: "du" / Catalan: "del"

### 4. Diacritics and Special Characters
For single diacritic characters or very short tokens:
- If it's just a character (like ò, è, ç), return it unchanged for both languages
- Example: ò → French: "ò" / Catalan: "ò"

### 5. Output Format
Return ONLY valid JSON with no additional text, markdown formatting, or code
```

```
blocks. The JSON must be an object where:
- Keys are the Occitan words (exactly as provided)
- Values are objects with "fr" (French) and "ca" (Catalan) keys

Example output for ["l'òme", "femna", "lo"]:
{"l'òme": {"fr": "l'homme", "ca": "l'home"}, "femna": {"fr": "femme", "ca": "dona"},
 "lo": {"fr": "le", "ca": "el"}}
```

The per-batch user prompt was:

```
Translate these Occitan words to French and Catalan.

Words to translate:
{json.dumps(tokens, ensure_ascii=False)}

Return ONLY the JSON object, no other text.
```

Representative generated alignment weights:

```
{
  "d'aquò": [["de cela", 0.198], ["d'això", 0.802]],
  "l'òme": [["l'homme", 0.484], ["l'home", 0.516]],
  "qu'es": [["qu'est", 0.623], ["que és", 0.377]],
  "del": [["du", 0.228], ["del", 0.772]],
  "òme": [["homme", 0.478], ["home", 0.522]],
  "femna": [["femme", 0.576], ["dona", 0.424]]
}
```

A.3 Targeted Dataset Snippets

A.3.1 HPLT Data Cleaning

The filtering removed Wikipedia edit tags, boilerplate validation comments, short lines and selected junk n-grams. Table A.1 shows short excerpts rather than full documents. The relevant cleaning constants are:

```
REGEX_WIKI_EDIT = re.compile(r'\[\s*modificar.*?\s*\]', re.IGNORECASE)

JUNK_PHRASES = [
    "modificar la font",
    "clicar sul ligam",
    "terminar lo procès de validacion",
    "vòstre comentari es a mand",
    "recebre per e",
```

Record ID	Raw HPLT Excerpt	Filtered Result Excerpt	Cleaning Signal
949f0...	... Origines e evolucion deth catalanisme politic[modificar | modificar la font] Per compréner ...	Eth catalanisme o nacionalisme catalan ei ua correnta sociau ... Origines e evolucion deth catalanisme politic Per compréner ...	REGEX_WIKI_EDIT deletes the bracketed edit control while retaining the surrounding article text.
95b31...	... I a pas cap de comentari Vòstre comentari es a mand d'èsser validat. Per terminar lo procès ...	Las resèrvas de l'arca de Noè Èra pas lo tot pel brave Noè de recampar las bèstias ...	Comment-validation boilerplate is removed by the junk-phrase filter.
f073a...	... Donc un renovelament e pas de far passar cossí foguesse lo prumier còp Vòstre comentari es a mand ...	Occitans e catalans s'amassan per 30n còp al pòrt de Salau Ongan, a l'ocasion del 30n anniversari ...	The article body is retained while reader-comment workflow text is dropped.

Table A.1: Comparison of raw HPLT web scrapes against the cleaned corpus excerpts.

```
    "anatz recebre per",
    "cal encara clicar",
    "dins lo navigador"
]
```

A.3.2 Tokeniser Stress-Testing Examples

A.3.3 Localised Localities and Pragmatic Norms

The localised outputs emphasize native Occitan naming (`Joan`), regional travel hubs (`Tolosa`, `Montpelhièr`, `Besièrs`), and everyday pragmatic norms.

Category	Example Sentence	Targeted Phenomenon
elision_overload	L'amic d'aquela femna qu'es aicí s'es trompat d'ostal.	Dense apostrophe/elision patterns: L', d', qu', s', d'.
elision_overload	N'i a pas qu'un que s'interèssa d'aquò e qu'o vòl far.	Multiple clitic and elided forms in a short sequence.
diacritic_density	Aquò's la leiçon de francés qu'estudièrem ièr al collègi.	ò, ç, é, è, and apostrophe handling in one sentence.
diacritic_density	L'òme diguèt que la clau èra amagada darrièr lo canapè.	Dense open-vowel accents and final è under normal sentence syntax.

Table A.2: Examples from the generated morphological stress-test corpus.

Source	Before / Scenario	Localised Occitan Output
Catalan Alpaca row index=4	source_input: En Joan li encanta tocar la guitarra. source_output: Tocar la guitarra és una activitat que li agrada molt a John.	input: Al Joan li agrada fòrça tocar la guitarra. output: Tocar la guitarra es una activitat que li agrada fòrça al Joan.
Catalan Alpaca row index=363	source_output: Genereu una història de ficció sobre un personatge anomenat "John"...	output: Generatz una istòria de ficcion sus un personatge nomenat 'Joan'...
Dialogue scenario: Buying train tickets	Station transaction, local travel context	Bonjorn, me cal un bilhet per Tolosa. N'i a encara? / Òc, n'i a. Vos lo cal per uèi o per deman?
Dialogue scenario: Asking family news	Family-news pragmatics and regional mobility	Ont es passada ta sòrre? Se ditz que demòra pus aicí. / Non, se n'es anada a Montpelhièr fa dos meses per i trabalhar.
Dialogue scenario: Asking bus schedules	Everyday transport norm	Quant ne còsta, d'anar fins a Besièrs? / Ne còsta pas gaire, se n'as la carta d'abonament.

Table A.3: Examples of cultural localisation and pragmatic alignment.

B Code Sample

B.1 InjectAuxLoss Autograd Module

This autograd function injects the Switch-Transformer-style load-balancing gradient directly into the backward pass. This was necessary because distributed and advanced training frameworks can sever or hide global auxiliary-loss accumulation paths when gradient checkpointing or sharding is active.

```
class InjectAuxLoss(torch.autograd.Function):

    @staticmethod
    def forward(ctx, gate_logits, num_experts, aux_loss_coef):
        ctx.save_for_backward(gate_logits)
        ctx.num_experts = num_experts
        ctx.aux_loss_coef = aux_loss_coef
        return gate_logits

    @staticmethod
    def backward(ctx, grad_output):
        gate_logits, = ctx.saved_tensors
        E = ctx.num_experts
        coef = ctx.aux_loss_coef

        if coef <= 0.0:
            return grad_output, None, None

        with torch.enable_grad():
            logits_tmp = gate_logits.detach().requires_grad_(True)
            gate_weights = F.softmax(logits_tmp, dim=-1)

            flat = gate_weights.reshape(-1, E)
            f_i = flat.mean(dim=0)

            L_bal = E * (f_i * f_i).sum()
            L_bal.backward()
```

```
            grad_aux = logits_tmp.grad

        # Add the auxiliary gradient to the task's gradient
        grad_input = grad_output + (coef * grad_aux)

        return grad_input, None, None
```

The injection point is located inside the MoLE LoRA forward replacement, immediately after router logits are computed and before softmax routing:

```
#  3. Auxiliary Loss Injection (FSDP Safe)
# During training, torch.is_grad_enabled() is True. We apply our
# autograd function to inject L_bal gradients.
if torch.is_grad_enabled() and per_layer_aux_coef > 0.0:
    gate_logits = InjectAuxLoss.apply(gate_logits, E, per_layer_aux_coef)
```

B.2 MoLEDDPTrainer Checkpoint Override

The final MoLE stage trains only router parameters over frozen LoRA experts. To avoid duplicating the full base language model at each checkpoint, the trainer overrides the Hugging Face `Trainer._save()` method and exports strictly the trainable router state alongside a compact JSON configuration.

```
class MoLEDDPTrainer(Trainer):
    """Trainer that saves only router weights/config."""

    def _save(self, output_dir: str | None = None, state_dict=None):
        if not self.args.should_save:
            return

        output_dir = output_dir or self.args.output_dir
        os.makedirs(output_dir, exist_ok=True)

        model = self.model
        if hasattr(model, "module"):
            model = model.module

        torch.save(
            model.routers.state_dict(),
            os.path.join(output_dir, "router_weights.pt"),
        )

        router_cfg = {
            "adapter_names": model.adapter_names,
```

```
            "num_experts": model.num_experts,
            "sequence_route_threshold": model.sequence_route_threshold,
            "router_aux_loss_coef": model.router_aux_loss_coef,
            "router_supervision_coef": getattr(model, "router_supervision_coef", 0.0),
            "router_supervision_class_weights": getattr(
                model,
                "router_supervision_class_weights",
                None,
            ).tolist() if hasattr(model, "router_supervision_class_weights") else None,
            "router_temperature": getattr(model, "router_temperature", 1.0),
            "hard_router_argmax": getattr(model, "hard_router_argmax", False),
            "router_modes": model.router_modes,
        }
        with open(os.path.join(output_dir, "router_config.json"), "w", encoding="utf-8")
            json.dump(router_cfg, f, indent=2)

        print(f"[MoLE/DDP] Saved router weights + config -> {output_dir}")
```

This override ensures that each checkpoint outputs exactly the routing layer weights and the metadata needed by downstream evaluation scripts to reconstruct the MoLE wrapper cleanly:

- `router_weights.pt`
- `router_config.json`