Adapting LLMs to Understand Vernacular Languages: A Case Study on Italian Varieties

Anonymous ACL submission

Abstract

Large language models perform strongly on standard languages but continue to struggle with vernacular varieties, which remain heavily underrepresented in the corpora and benchmarks used to train modern systems. These varieties are often primarily spoken and lack stable codified grammars, making the creation of high-quality annotated resources particularly challenging. As a result, substantial performance gaps persist for vernacular input in low-resource settings. To address these challenges, we introduce VERNADAPT, a framework for bootstrapping instruction-tuning data for vernacular language understanding through data distillation from raw vernacular sources while grounding outputs in the corresponding standard language. By focusing on *understanding* rather than free-form generation, the framework avoids requiring fully standardized vernacular outputs and limits native-speaker involvement to test set validation, substantially reducing annotation costs. We focus on Italian vernacular varieties as a particularly diverse linguistic setting and use VERNADAPT to build LAVITA, a parallel multitask dataset covering three varieties, and fine-tune several multilingual and monolingual models. Our results show that multilingual pretrained and instruction-tuned models consistently outperform monolingual ones, particularly when the target variety is linguistically close to the standard language, and that joint adaptation across varieties often matches or exceeds single-variety training.

1 Introduction

Despite remarkable progress on standard languages, modern Large Language Models (LLMs) continue to struggle with the hundreds of vernacular varieties spoken daily by millions of people worldwide (Faisal et al., 2024). This performance degradation has been documented across tasks and language families, affecting even the most advanced frontier models (Lin et al., 2025). These shortcomings likely arise from two compounding factors: vernacular varieties are often linguistically distant from their standardized counterparts, limiting transfer from existing models, and they remain systematically underrepresented in both training corpora and evaluation benchmarks (Plank, 2016; Lutgen et al., 2026).

However, non-standard varieties[1] are crucial vectors of cultural knowledge, literary works, and community histories that often do not exist in standard languages. While they have traditionally been oral, they now exhibit a large and growing digital footprint through user-generated content such as social media posts, song lyrics, subtitles, and online forums, reflecting their vitality and enabling their systematic study in NLP (Blaschke et al., 2023; Joshi et al., 2024). Yet speakers of these varieties already interact daily with AI systems (e.g., virtual assistants and LLM-powered search engines) that were not built with their languages in mind (Blaschke et al., 2024), reflecting a broader Western- and Chinese-centric bias in LLM development and raising the question of *who* gets to benefit from progress in language technology (Blasi et al., 2022).

Italian vernacular languages are a particularly striking case of this neglect. Italy has one of the richest linguistic diversity landscapes in Europe; varieties like Roman, Neapolitan, and Sicilian alone count millions of speakers and sustain an active presence in music, cinema, and social media, alongside a deep literary heritage (Ramponi, 2024). Yet NLP research for Italian varieties remains scarce, and most are at risk of disappearing within a few generations (Moseley, 2010; Ramponi, 2024). Even the most capable models show notable performance gaps on vernacular input, and closing this gap through conventional training is

[1] In this work, we use the terms *vernacular* and *non-standard* interchangeably to refer to language varieties that diverge from standardized written norms.

prohibitively expensive, requiring native speakers to produce and curate data at a scale that cannot realistically extend to dozens of varieties (Sprugnoli et al., 2025; Signoroni and Rychlý, 2026).

The goal of this paper is to show that, in the absence of a stable written standard, understanding of vernacular varieties can be enabled by synthetic, supervised fine-tuning (SFT) data, thereby reducing the need for costly manual annotation. To this end, we put forward VERNADAPT, a framework for the automatic generation and validation of SFT data: The examples are constructed from gold vernacular source texts, while all instructions and target responses are generated in the associated standard language, requiring native speaker involvement only for test set validation. This design effectively decouples the ability to *understand* a vernacular variety from the ability to *generate* it.

We apply our framework focusing on three Italian varieties, building LAVITA, a parallel multitask SFT dataset which we use to fine-tune several models, both individually and jointly. We further investigate whether downstream performance correlates with the linguistic distance between a given variety and the standard languages seen during training.

Our results show that the gap between vernacular and standard Italian understanding is substantial and consistent across tasks, but also partially bridgeable through modest fine-tuning on synthetic data of reasonable quality. We additionally find that models with broad multilingual pretraining and Italian instruction tuning achieve the strongest performance, and that varieties that are linguistically closer to standard Italian benefit more from adaptation. We release our code and data at `omitted.link` under CC BY-SA 4.0.

2 Related Work

NLP for vernacular varieties The study of language varieties in NLP has gained increasing attention with dedicated venues such as VarDial (Scherrer et al., 2023) and recent surveys (Blaschke et al., 2023; Joshi et al., 2024). Studies on English varieties show consistent performance degradation on non-standard input (Lin et al., 2025; Ziems et al., 2022). Multilingual benchmarks confirm this pattern across language families (Faisal et al., 2024), with LLMs producing lower-quality and more stereotyped outputs for non-standard varieties (Fleisig et al., 2024).

In contrast to English and other high-resource multilingual settings, NLP for Italian vernacular varieties remains sparse and fragmented. Existing efforts are primarily resource-driven and include parallel and annotated corpora for individual varieties (Haberland et al., 2024; Alam et al., 2024; Cappello et al., 2025), as well as lexical resources for structured knowledge bases (Sprugnoli et al., 2025). Beyond resource creation, recent work has begun to evaluate LLM behaviour on Italian varieties (Lilli, 2023) and to explore fine-tuning for vernacular conversational systems (Pistagnesi, 2024), but systematic adaptation of LLMs to Italian vernacular varieties remains largely unexplored.

Synthetic data for instruction tuning. Synthetic data generation has become central to instruction tuning, enabling supervision at scale without extensive human annotation. Self-Instruct (Wang et al., 2023) showed that model-generated pairs can effectively align smaller models, while Zhou et al. (2023) demonstrated that quality and composition matter more than scale. Synthetic data can be effective even in low-resource settings, particularly when existing corpora are transformed into instruction-following examples (through operations such as question generation or summarization) and generation is carefully controlled (Pengpun et al., 2024; Anikina et al., 2025).

Our approach follows this task-grounded perspective but differs in two ways: we construct supervision directly from authentic vernacular texts rather than curated or generic corpora, and we restrict generation to a small set of core NLU tasks allowing tighter control over quality and task distribution in a setting where both linguistic variation and resource scarcity pose additional challenges.

Tokenizer fertility and cross-lingual performance Tokenizer fertility (i.e. the average number of tokens per word) is a standard measure of tokenizer coverage for a given variety (Petrov et al., 2023; Ali et al., 2024b). High fertility increases sequence length and, consequently, computational cost, with documented overheads of up to 68% for non-primary languages (Ali et al., 2024b). Fertility has also been shown to correlate with downstream performance (Ali et al., 2024b; Azime et al., 2025), although the relationship is task-dependent. We use fertility alongside linguistic distance as a proxy for cross-lingual transfer potential.

PEFT for low-resource settings. Parameter-Efficient Fine-Tuning (PEFT) methods such as

LoRA enable Large Language Model adaptation with minimal parameters, making them well-suited for low-resource scenarios. Recent work has extended PEFT to vernacular adaptation via variety-specific adapters (Liu et al., 2023), though primarily for English varieties or in relatively resource-rich settings. The works most closely related to ours are Chatzikyriakidis et al. (2026) and Khan et al. (2026), which use LoRA-based continual training to adapt models to *produce* Greek varieties and Québec French, respectively, demonstrating consistent gains even in low-resource conditions. In contrast, we target vernacular *understanding* via supervised instruction tuning with synthetic data, grounding model outputs in the standard language to enable straightforward evaluation.

3 VERNADAPT: A Framework for Vernacular Language Understanding

We propose VERNADAPT, a lightweight framework for SFT dataset construction targeting vernacular language *understanding* in settings where the target variety lacks a stable written standard. Such settings pose two distinct challenges: (i) the absence of a standardized grammar makes training a model to *produce* the target variety an ill-posed problem, as there is no well-defined target form to learn; (ii) annotated training data is scarce or entirely absent. We address both challenges in VERNADAPT: the first by grounding all model outputs in the standard language, thereby making the task well-posed and evaluation tractable; the second by generating SFT samples synthetically from raw vernacular source texts, requiring only minimal native speaker involvement.

We test VERNADAPT on Italian vernacular varieties, whose breadth and diversity make them one of Europe's most compelling linguistic testbeds (Ramponi, 2024), selecting Roman, (ROM) Neapolitan (NAP), and Sicilian (SCN) as three varieties at increasingly distant points from the standard. In the following, we describe our framework and the resulting dataset, LAVITA.

3.1 Source Data and Variety Selection

We select three Italian varieties based on the availability of written resources, and degree of divergence from standard Italian. The three varieties represent distinct systems that diverge from the standard (Figure 1). While Roman has historically converged closely with standard Italian, Neapolitan and Sicilian had a more independent development due to external sociodemographic factors (Appendix A).

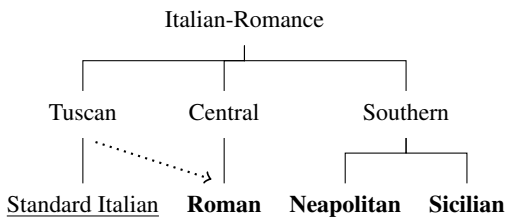

Figure 1: Lineage of the analyzed Italian-Romance varieties. The solid lines represent genetic descent, while the dotted arrow maps the inter-variety convergence (*Tuscanization*) of Roman toward Italian.

All data was gathered from copyright-free web sources. We distinguish two domains: *Book*, comprising literary and narrative prose drawn from LiberLiber[2] and Wikisource[3], available for all three varieties; and *Wiki*, comprising encyclopedic prose from the Wikipedia projects, available for Neapolitan and Sicilian only.

3.2 Task Design

A core design decision in VERNADAPT is to restrict evaluation to *understanding* tasks, excluding open-ended vernacular generation. As discussed above, non-standard Italian varieties lack a codified grammar, making generation an ill-posed objective: there is no well-defined target form against which model outputs can be reliably scored. By contrast, comprehension tasks can be evaluated against standard-language reference outputs, yielding tractable and reproducible results. Thus, we select three understanding tasks that probe different aspects of language understanding at increasing levels of difficulty and text-level granularity, namely question answering (QA), summarization, and translation (cf. Table 15). Throughout this work, we use SL to refer to *standard language* (e.g., Italian) and VL to refer to *vernacular language* (e.g., Sicilian).

Question answering (QA) Given a VL passage and SL question, the model must produce an SL answer. QA targets *local* comprehension: the model must identify and interpret a specific span of the source text, but need not construct a global representation of its content. It is the most constrained

[2] https://liberliber.it/
[3] https://it.wikisource.org/

of the three tasks, with answers typically short and verifiable.

Summarization Given a VL passage and a SL summarization instruction, the model must produce a SL summary. Summarization targets *global* comprehension: unlike QA, the model must integrate information across the full passage and abstract away from surface form to produce a coherent output. This makes it more demanding than QA, requiring a deeper and broader understanding of the source text.

Translation (MT) Given a VL passage, the model must produce a SL translation. Translation is the most demanding of the three tasks: it requires both global comprehension of the source text and fine-grained knowledge of variety-specific lexis, morphology, and syntax. Unlike the previous tasks, it also requires full coverage of the source content, leaving no room for selective understanding.

3.3 Dataset Construction

We begin by segmenting the source documents into text chunks, which serve as the basic unit for generating SFT items. We set a fixed minimum chunk size of 150 words while preserving sentence boundaries. We then aggregate the resulting chunks so that their distribution matches that of the original source documents, in order to mitigate potential length biases. Domain proportions are preserved via stratified sampling: for Neapolitan and Sicilian, approximately 20% of chunks are drawn from the Book domain with the remainder coming from Wiki source; for Roman, all chunks come from the Book domain, as no Wiki source is available (Section 3.1). Following Zhou et al. (2023), we select 1,000 chunks per variety for training, and 100 each for validation and test.[4].

Synthetic SFT generation SFT samples are generated synthetically from the source chunks using Gemini-2.5-Pro (Comanici et al., 2025), selected on the basis of three criteria: its strong performance on Sicilian-to-Italian translation (Sprugnoli et al., 2025), its demonstrated effectiveness for synthetic data generation in Italian (Gioffré et al., 2026), and its superior output quality in our own preliminary exploration across candidate models.

For training data, each chunk is used to generate a single task instance, with tasks assigned as follows: 50% of chunks are used for QA, 25% for summarization, and 25% for MT.[5] For validation and test data, each chunk is used to generate instances for all three tasks. Due to the smaller size of source data for Roman, each training chunk is used to generate instances for all three tasks.

For each task, a chunk yields the following outputs: two question-answer pairs in Italian for QA; a short Italian summary for summarization; and a full Italian translation for MT. Instructions are randomly sampled from a set of task-specific templates (see Table 17) and their position relative to the source chunk is randomized, to reduce prompt sensitivity in evaluation. Summarization and MT items are always paired with a randomly sampled task instruction; for QA, an explicit instruction is included only in 50% of instances, as bare question prompts are also suitable for this task.

Native speaker validation While we use training and validation instances *as-is*, all test items undergo human validation. For each variety, we recruited three native speakers, selected to reflect intra-variety demographic variation in age, geographic provenance, and self-reported vernacular proficiency (Table 7). Each annotator was asked to correct errors in the generated outputs and provide a justification when the correction was not self-evident (Table 16). Annotators frequently flagged different words or spans as incorrect, as expected given the demographic diversity of our annotator pool. All corrections were then manually reviewed and integrated, yielding a high-quality test set.

Qualitative error annotation We also asked a trained linguistic annotator to categorize all native speaker corrections by error type and severity, according to a set of guidelines (Appendix B). Our analysis reveals that Gemini exhibited relatively high item-level error rates across conditions, but substantially lower word-level error rates (Table 1). This suggests that, although many items contained at least one error, the errors themselves were typically localized to only a few words and did not substantially alter the meaning or structure of the original text in many cases (see Appendix E for error breakdown).

3.4 Dataset Statistics

Table 2 reports the size of the resulting dataset, LAVITA, across varieties, domains, and splits. For

[4]We split the data at the document level to avoid content contamination across sets.

[5]We use individual chunks for creating MT task instances.

Variety	Error rate	
	Item-level	Word-level
NAP	0.3100	0.0117
SCN	0.2225	0.0217
ROM	0.1798	0.0149
Avg.	0.1351	0.0161

Table 1: Average Gemini item- and word-level error rates when generating the SFT items.

Variety	Domain	Train	Val	Test
NAP	Book	305	175	69
	Wiki	1213	370	293
SCN	Book	269	86	71
	Wiki	1279	345	292
ROM	Book	1400	365	351
Total	All	4466	1341	1076

Table 2: SFT instances by variety, domain, and split.

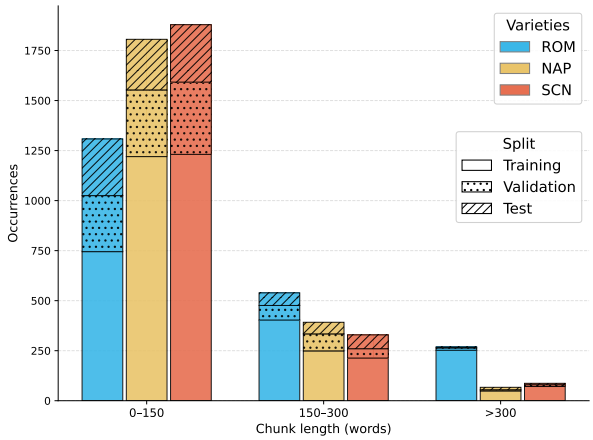

Figure 2: Chunk length distribution across varieties.

each of the three varieties, the dataset contains approximately 1500 training instances, 400 validation instances, and 350 test instances, for a total of 6883 instances overall. We report the distribution of chunk lengths across the three varieties in the dataset in Figure 2.

4 Experimental Setup

Models A central question of this work is whether a model's exposure to related Romance languages during training translates into measurable gains after adaptation to Italian varieties. This drives our selection of a diverse set of instruction-tuned models grouped by their training languages (training data composition in Tables 18 and 19).

We define a model as *Italian-specialized* if it was trained on at least 30% Italian tokens during pretraining, or if it underwent continual pretraining or supervised fine-tuning on a majority of Italian tokens or examples. These include three variants of Minerva (Orlando et al., 2024; Moroni et al., 2025a), a family pretrained from scratch on Italian data (v1, v1.5-Recipe1, and v1.5-Recipe2), and four models adapted from existing multilingual backbones: ANITA (Basile et al., 2023; Polignano et al., 2024), Llama-LAPT and Llama-SAVA (Moroni et al., 2025b) – three Italian adaptations of Llama 3 (Grattafiori et al., 2024) differing in adaptation strategy (instruction tuning, language-adaptive pretraining, and vocabulary adaptation respectively) – and Occiglot-it (Avramidis et al., 2024), a Mistral-based (Jiang et al., 2023) model instruction tuned on Italian and English.

We consider as *Romance-specialized* those models trained primarily on at least one Romance language other than Italian, either constituting a substantial portion of pretraining data or through continual pretraining or fine-tuning on such a language. This group includes Lucie (Gouvert et al., 2025) and the -es, -fr, and -eu5 Occiglot variants, instruction-tuned on Spanish, French and the five major European languages respectively (Avramidis et al., 2024).[6]

We group in *European-specialized* all those models, such as EuroLLM (Ramos et al., 2026), Salamandra (Gonzalez-Agirre et al., 2025), and Teuken (Ali et al., 2024a), that were optimized across the full spectrum of European languages, which constituted the majority of their pretraining or adaptation data.

Finally, we distinguish two further groups. *English-specialized* models are those trained primarily on English with no particular focus on other languages; OLMo3 (Olmo et al., 2026) is our sole representative of this category, selected for its fully transparent training data composition, which makes it a clean baseline. *Broad-coverage* models such as Llama-3.1 (Grattafiori et al., 2024), Gemma 2 (Gemma, 2024), and Qwen3 (Qwen, 2025) are general-purpose systems whose training data composition is not publicly disclosed in sufficient detail for the purposes of our analysis, but we assume it to be diverse enough, due to their proficiency in several languages, serving as baselines with no known non-English specialization.

[6] All Occiglot models share a continual training phase covering English, French, German, Italian, and Spanish.

ID	Rank	LR	Epochs	Freq.
M4	16	3×10^{-4}	3	12
M3	16	1×10^{-4}	3	8
L1	32	1×10^{-4}	3	8
L8	32	3×10^{-4}	5	7
M5	16	4×10^{-4}	3	7

Table 3: LoRA configurations appearing in the top-5 training runs across dialect adapters, selected via rank aggregation, ordered by their frequency.

Task	Dim.	METEOR	GPT-oss-120b
QA	Factuality	0.606 [0.49, 0.70]	0.773 [0.65, 0.86]
Sum.	Faithful.	0.363 [0.21, 0.51]	0.522 [0.40, 0.63]
	Coverage	0.507 [0.38, 0.62]	0.625 [0.50, 0.73]

Table 4: Correlation between automatic evaluation metrics and human judgments across different dimensions. We report Kendall's τ with 95% bootstrap CI.

Control settings We compare the results of the adapted models against four control settings across tasks. The *zero-shot* setting serves as the primary control, where each model receives the task prompt with no demonstrations. In contrast, the *few-shot* setting provides three demonstrations per task and variety, where examples consist of vernacular source text with Italian instructions and responses. We also define *Italian-context setting* for both tasks, in which the model receives the SL version of the source text[7] establishing the ceiling performance on standard Italian input. For QA, we additionally include a *question-only setting* in which the model is queried with only the question to probe parametric knowledge in isolation.

LoRA fine-tuning We train one LoRA adapter per variety plus a joint adapter trained on the full combined dataset. For Italian-specialized models, adapters were trained across a grid of hyperparameters, resulting in 16 configurations per variety (Table 8). To reduce computational cost while maintaining coverage, non-Italian-specialized models were fine-tuned only on the top-5 configurations per variety, selected by downstream performance on Italian-specialized models via rank aggregation. Each configuration was ranked per task (rank 1 being best), and the configuration with the lowest rank sum across the three tasks was chosen as the final adapter. Table 3 reports the resulting best hyperparameters for each adapter.

Evaluation metrics Metrics are selected on the basis of their correlation with human judgement for open-ended generation tasks. Drawing on Bonomo et al. (2025), who identify METEOR and LLM-as-a-Judge as the most reliable metrics for open-ended QA, we adopt both metrics as a starting point and verify their reliability on our dataset by computing Kendall's τ with bootstrap against human judgements on an annotated subset of our test set (Deutsch et al., 2022). For each task, we randomly sample 25 items from the test set, collecting outputs from one model per item chosen at random. These outputs are sampled with stratification across zero-shot, upper baseline, and fine-tuned settings, to ensure as much diversity as possible. We extend this correlation analysis to the summarization task as well (Table 4).

We evaluate QA on factual correctness, while we use faithfulness to the source and coverage of key information for summarization. In both cases, we use GPT-oss-120b (OpenAI, 2025) as the judge model. We prompt the judge using the same guidelines adopted to perform the evaluation in our experiments, and report them in Figures 13 to 15.

For MT, we adopt xCOMET (Guerreiro et al., 2024) in the reference-based setting, scoring hypotheses against a reference translation without relying on the source.

Datasets We train and evaluate all models on our dataset, LAVITA. For the translation task, we also employ another dataset, CODET (Alam et al., 2024), which consists of parallel sentences (VL and SL) for many geographical region/city associated with one variety.

5 Results

5.1 Aggregated Results by Model Group

Figures 3 to 5 show the performance of our tested models on QA, summarization, and MT across all experimental settings aggregated by the language groups we previously defined. We report additional performance on summarization and MT in Appendix D for space constraints.

Variety-level pattern A consistent performance trend emerges across all tasks and settings: ROM outperforms NAP, which in turn outperforms SCN. This hierarchy holds in zero-shot, few-shot, LoRA, and joint LoRA conditions alike, suggesting it reflects stable properties of the varieties rather than

[7] All chunks are translated as a by-product of our pipeline.

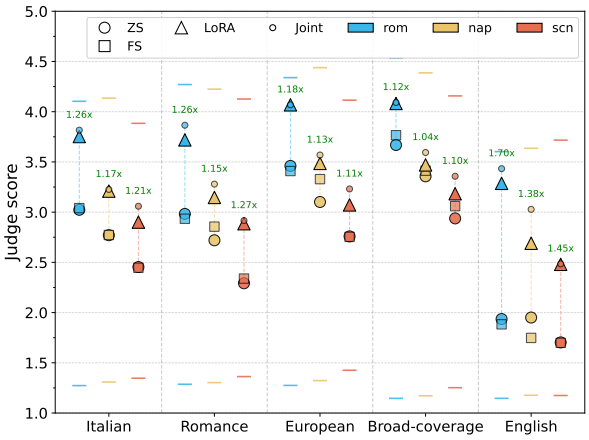

Figure 3: QA performance grouped by our model classes. We show the increment of LoRA finetuned models with respect to the zero-shot baseline. Horizontal lines are the two control settings: question-only (lower) and Italian-context (upper).

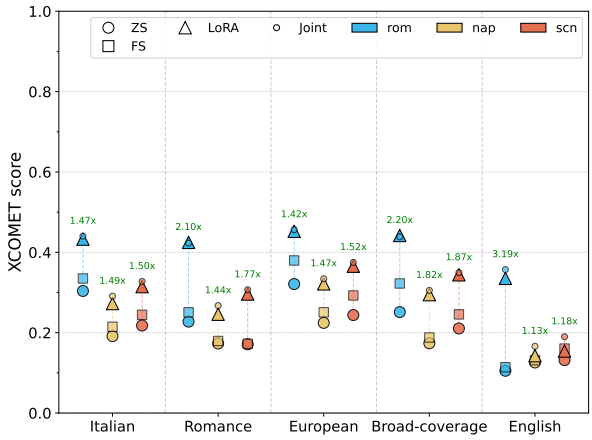

Figure 5: MT performance scored by XCOMET grouped by our model classes. We show the increment of LoRA finetuned models with respect to the zero-shot baseline.

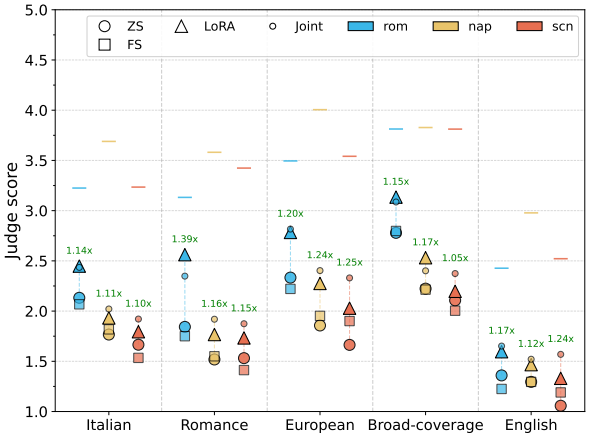

Figure 4: Summarization performance (Faithfulness) grouped by our model classes. We show the increment of LoRA finetuned models with respect to the zero-shot baseline. Horizontal lines are the Italian-context setting.

condition-specific artifacts. We attribute this pattern to two reinforcing factors. First, ROM is linguistically closer to standard Italian than NAP and SCN, which plausibly reduces the distributional shift the model must overcome. Second, our SCN data was contributed by annotators from across the island, resulting in greater orthographic and lexical variability; annotators frequently disagreed on unfamiliar terms, whereas ROM and NAP texts, drawn from geographically more compact communities, were handled more consistently. These two factors are partially confounded, but both predict the observed ordering.

MT ranking Translation follows a slightly different vernacular hierarchy: ROM > SCN > NAP, both on our primary dataset and on CODET (Figure 12). The inversion of NAP and SCN relative to QA and summarization is notable. Our dialectometric analysis through IRI[8] provides supporting signal: when computed over the full grammar cues, the similarity ordering to Italian is ROM > SCN > NAP, mirroring the MT results; when restricted to lexical cues only, the ordering shifts to ROM > NAP > SCN, aligning instead with QA and summarization. This suggests that translation is more sensitive to overall vernacular similarity, capturing phonological and morphological features alongside lexical variation, whereas comprehension tasks are driven more specifically by lexical distance.

Effect of model group Romance- and European-specialized models achieve the highest absolute scores, followed by Italian-specific models, with English-only models trailing. This ordering reflects the degree to which models have been exposed to related languages during pretraining: proximity to Italian in the training distribution provides a non-trivial head start even before adaptation. Notably, LoRA adaptation produces large relative gains for English-only models, confirming that the adapters are injecting genuine vernacular knowledge rather than merely amplifying existing competence (Appendix E) – yet absolute performance remains lower, suggesting that parametric knowledge of related languages cannot be fully compensated by fine-tuning alone. Closed-source models score competitively, although their training data compo-

[8] IRI is quantitative metric that measures the dialectometric distance between geographic locations (Goebl, 1989).

Variety	IRI (all)	IRI (only lexical)
NAP	0.64	0.59
SCN	0.65	0.55
ROM	0.80	0.73
ITA	1.00	1.00

Table 5: IRI similarity scores with respect to Standard Italian by variety (higher is more similar).

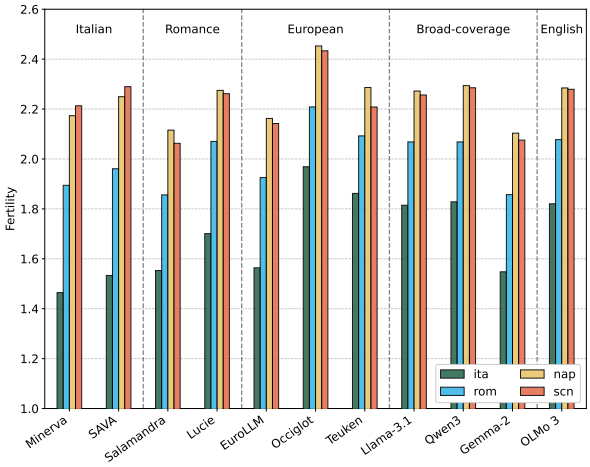

Figure 6: Models' fertility over LAVITA's test set.

sition is unknown, precluding direct attribution.

5.2 Model-specific Results

Training data and performance We report the results of individual models in Table 14. The top five models overall are Gemma, Llama-LAPT, EuroLLM, ANITA, and Occiglot eu5, followed closely by Occiglot it and Salamandra. Across all tasks and varieties, the strongest predictor of zero-shot performance is the combination of broad multilingual pretraining covering Romance languages and Italian-specific instruction tuning; models satisfying both conditions consistently rank among the best. Although Gemma is a closed-source model and its training data composition is unknown, its strong performance suggests it likely falls into this category as well. Salamandra, despite lacking Italian-specific instruction tuning, frequently appears in the top seven, suggesting that pretraining on related languages still provides a useful foundation, even when Italian is not seen during post-training. By contrast, Teuken and Qwen, despite being multilingual LLMs, do not reach comparable performance levels, likely because their pretraining data is less focused on Italian and Romance languages specifically.

Tokenizer fertility and linguistic distance Fertility scores, reported in Figure 6, correlate with linguistic distance from standard Italian, both from a genealogical standpoint (Figure 1) and according to the IRI dialectometric distance (Table 5). The ordering ITA $<$ ROM $<$ SCN $\approx$ NAP reflects increasing distance from Italian, with NAP and SCN showing comparable and consistently higher fertility than ROM, though their relative ordering occasionally varies. This alignment suggests that fertility captures something meaningful about the structural divergence of each variety from the model's primary training language.

However, fertility scores do not straightforwardly predict downstream performance. The Occiglot family, which exhibits the highest fertility across all varieties, consistently ranks among the top-performing models. This suggests that subword fragmentation of dialectal text is not the only bottleneck in vernacular adaptation; factors such as pretraining language coverage and instruction tuning are stronger predictors of performance.

6 Conclusion

We presented VERNADAPT, a framework for generating synthetic SFT data for vernacular language understanding, and released LAVITA, a multitask dataset covering Roman, Neapolitan, and Sicilian across QA, summarization, and MT. Our experiments show that LoRA adaptation on synthetic vernacular data consistently improves over zero-shot and few-shot baselines, demonstrating that substantial gains are achievable without large manually curated resources. Importantly, VERNADAPT sidesteps the lack of standardized orthographies common to non-standard varieties by not requiring models to generate vernacular output.

Our analysis reveals that tokenizer fertility, while correlated with linguistic distance from standard Italian, is not by itself a reliable predictor of downstream performance after adaptation. Instead, the strongest predictor of both zero-shot and adapted performance is the combination of broad multilingual pretraining – particularly on related language families – and instruction tuning in the standard language, with varieties closer to the standard consistently benefiting more from transfer. These findings offer practical guidance for researchers approaching vernacular adaptation in other linguistic contexts.

Limitations

The primary linguistic challenges include inconsistent orthographic conventions and substantial intra-dialectal variation, particularly in Sicilian given its broader geographic distribution. We partially mitigate this by framing tasks around language *understanding* rather than *generation*, sidestepping the need for orthographic normalization. Our scope is further limited to three southern-central varieties; while Italian vernacular languages constitute one of Europe's richest linguistic landscapes, we leave to future work the analysis of how findings generalize across this broader range.

The synthetic training data is not manually verified at the same level as the test set, and future work could incorporate quality filtering to remove inconsistent or hallucinated dialectal forms. Task coverage is intentionally narrow — QA, summarization, and translation — as a proof of concept for effective LoRA adaptation on vernacular varieties rather than a comprehensive fine-tuning effort. Finally, we do not perform continual pre-training on dialectal text, which could inject broader morphosyntactic knowledge of the varieties, albeit complicating evaluation.

References

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2024. CODET: A benchmark for contrastive dialectal evaluation of machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, and 20 others. 2024a. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *Preprint*, arXiv:2410.03730.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024b. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Tatiana Anikina, Jan Cegin, Jakub Simko, and Simon Ostermann. 2025. A rigorous evaluation of LLM data generation strategies for low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8282–8303, Suzhou, China. Association for Computational Linguistics.

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, and Kristian Kersting. 2024. Occiglot at wmt24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*. Association for Computational Linguistics.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6265–6281, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *Preprint*, arXiv:2312.09993.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? a survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tommaso Bonomo, Luca Gioffré, and Roberto Navigli. 2025. LiteraryQA: Towards effective evaluation of long-document narrative QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, pages 34086–34107, Suzhou, China. Association for Computational Linguistics.

Caterina Maria Cappello, Sabrina D'Alì, Mario Guglielmetti, Elisa Di Nuovo, and Cristina Bosco. 2025. Arbuli sunnu: A Sicilian-Italian parallel treebank. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 155–168, Cagliari, Italy. CEUR Workshop Proceedings.

Stergios Chatzikyriakidis, Dimitris Papadakis, Sevasti-Ioanna Papaioannou, and Erofili Psaltaki. 2026. Grdd+: An extended greek dialectal dataset with cross-architecture fine-tuning evaluation. *Preprint*, arXiv:2511.03772.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Team Gemma. 2024. Gemma.

Luca Gioffré, Luca Moroni, Alberte Fernández-Castro, Elena Marafatto, Giacomo Garufi, and Roberto Navigli. 2026. INDAQA2: A large Italian narrative QA benchmark. In *EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, number 4195 in CEUR Workshop Proceedings, Bari, Italy.

H Goebl. 1989. matrice de similarité en appliquant l'indice relatif d'identité (iri). In *New Methods in Dialectology: Proceedings of a Workshop Held at the Free University, Amsterdam, December 7-10, 1987*, volume 33, page 165. Walter de Gruyter.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Iñigo Pikabea, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 4 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Olivier Gouvert, Julie Hunter, Jérôme Louradour, Christophe Cerisara, Evan Dufraisse, Yaya Sy, Laura Rivière, Jean-Pierre Lorré, and OpenLLM-France community. 2025. The lucie-7b llm and the lucie training dataset: Open resources for multilingual language generation. *Preprint*, arXiv:2503.12294.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian machine translation in its cultural context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *Preprint*, arXiv:2401.05632.

Eeham Khan, Firas Saidani, Owen Van Esbroeck, Richard Khoury, and Leila Kosseim. 2026. Low-resource dialect adaptation of large language models: A french dialect case-study. *Preprint*, arXiv:2510.22747.

Silvia Lilli. 2023. Chatgpt-4 and italian dialects: Assessing linguistic competence. *Umanistica Digitale*, 7(16):235–263.

Fangru Lin, Shaoguang Mao, Emanuele La Malfa, Valentin Hofmann, Adrian de Wynter, Xun Wang, Si-Qing Chen, Michael J. Wooldridge, Janet B. Pierrehumbert, and Furu Wei. 2025. Assessing dialect fairness and robustness of large language models in reasoning tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6317–6342, Vienna, Austria. Association for Computational Linguistics.

Yanchen Liu, William Held, and Diyi Yang. 2023. DADA: Dialect adaptation via dynamic aggregation of linguistic rules. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13776–13793, Singapore. Association for Computational Linguistics.

Michele Loporcaro. 2013. *Profilo linguistico dei dialetti italiani*. LaTerza.

Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in nlp. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11332–11344, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Martin Maiden and M. Mair Parry, editors. 1997. *The Dialects of Italy*, illustrated edition. Romance Linguistics. Psychology Press.

Luca Moroni, Tommaso Bonomo, Luca Gioffré, Lu Xu, Domenico Fedele, Leonardo Colosi, Andrei Stefan Bejgu, Alessandro Scirè, and Roberto Navigli. 2025a. What we learned from continually training minerva: A case study on Italian. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 760–774, Cagliari, Italy. CEUR Workshop Proceedings.

Luca Moroni, Giovanni Puccetti, Pere-Lluís Huguet Cabot, Andrei Stefan Bejgu, Alessio Miaschi, Edoardo Barba, Felice Dell'Orletta, Andrea Esuli, and Roberto Navigli. 2025b. Optimizing LLMs for Italian: Reducing token fertility and enhancing efficiency through vocabulary adaptation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6661–6675, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 445–464, Bangkok, Thailand. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc.

Laura Pistagnesi. 2024. Fine-tuning a large language model to define a chatbot capable of operating in a marche dialect. Master's thesis, Università Politecnica delle Marche.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. In *KONVENS*.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

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

Miguel Moura Ramos, Duarte M. Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M. Guerreiro, Ricardo Rei, Nicolas Boizard, Amin Farajian, Mateusz Klimaszewski, José G. C. de Souza, Barry Haddow, François Yvon, Pierre Colombo, Alexandra Birch, and André F. T. Martins. 2026. Eurollm-22b: Technical report. *Preprint*, arXiv:2602.05879.

Alan Ramponi. 2024. Language varieties of italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19 to 38.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri, editors. 2023. *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*. Association for Computational Linguistics, Dubrovnik, Croatia.

Edoardo Signoroni and Pavel Rychlý. 2026. Can LLMs translate Italy's language varieties? In *Proceedings for the Ninth Workshop on Technologies for Machine Translation of Low Resource Languages (LoResMT 2026)*, pages 69–77, Rabat, Morocco. Association for Computational Linguistics.

Rachele Sprugnoli, Giovanni Moretti, Domenico Giuseppe Muscianisi, and Eleonora Litta. 2025. Ciallabacialla! modeling and linking a regional lexical resource to include Sicilian in the semantic web. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 1093–1101, Cagliari, Italy. CEUR Workshop Proceedings.

Pietro Trifone and 1 others. 2008. *Storia linguistica di Roma*. Carocci Editore.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Chunting Zhou, 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*, volume 36, pages 55006–55021. Curran Associates, Inc.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Italian varieties

In this Section, we describe and contextualize the three varieties used in this paper – Roman, Neapolitan, and Sicilian.

Roman (*Romanesco*) A central Italian variety spoken in Rome and its surrounding area, historically derived from 14th-century Tuscan varieties, the same underlying modern standard Italian, resulting in a variety that remains largely intelligible to standard Italian speakers. Unlike other Italian regional varieties, Roman began exhibiting lexical convergence with Tuscan as early as the 15th century, a process driven by significant sociohistorical factors. The presence of Tuscan papal courts in the 1400s fostered early linguistic influence, while a double demographic shock following the Sack of Rome in 1527 accelerated this shift: the population collapse among native Romans, combined with a subsequent doubling of the city's population through an influx of Tuscan merchants and settlers, reshaped the variety's lexical character well before similar processes took hold elsewhere in the peninsula. (Trifone et al., 2008)

Neapolitan (*Napoletano)* The Italo-Dalmatian variety of Naples and the surrounding Campanian municipalities, Neapolitan diverges substantially from standard Italian, exhibiting distinct phonology, morphosyntax, and a significant lexical stratum inherited from prolonged contact with Spanish, Greek, Catalan, and French.

Sicilian (*Sicilianu*) Spoken across the island of Sicily, Sicilian is an Italo-Dalmatian variety derived from Latin with substantial lexical influence from Arabic, Spanish, Greek, and Catalan, reflecting the island's layered history of contact (Maiden and Parry, 1997; Loporcaro, 2013). It is the most divergent of the three varieties, sitting beyond the principal isogloss bundles that separate southern from central Italo-Romance. Sicilian also boasts one of the oldest literary traditions among Italian varieties, with the 13th-century Sicilian School representing a foundational moment in the history of Italian poetry. Its Wikipedia edition (*Wikipèdia Siciliana*) constitutes one of the largest digitally available corpora for any Italian variety, making it a particularly tractable target for data collection.

B Annotation guidelines

B.1 Native speaker annotation

Table 16 shows the guidelines provided to the native speaker for validating and correcting LAVITA test sets (Section 3.3). We contacted native speakers through social media posts and each of them voluntarily opted to participate in the annotation process without compensation.

B.2 Error annotation guidelines

Table 6 contains the guidelines used to annotate model responses before and after fine-tuning. Labels are assigned along three independent dimensions: Language, Meaning, and Generation. Note that the label set is non-exhaustive and was defined a priori; annotators were free to introduce additional labels as needed. All tasks require responses in standard Italian.

Language dimension This dimension assesses whether LoRA fine-tuning corrupts Italian output, evaluated at the lexical and grammatical level. We distinguish between minor and major vernacular leakage. Minor leakage include *Lexical* and *Grammatical leakage*: the former is assigned when the response contains vernacular or dialect term for which a standard Italian equivalent exists, while the latter indicates the presence of vernacular grammatical structures. Instead, a major leakage happens when the response is written fully or predominantly in dialect (*Vernacular output*) or when it contains grammatical issues unrelated to vernacular influence (*Language error*).

Meaning dimension This dimension evaluates whether fine-tuning introduces semantic degradation. For QA and summarization tasks, we define *Hallucination* when the response contains invented, unsupported, or factually incorrect information. For translation, we label as *Translation error* those inaccuracies that alter or significantly distort the intended meaning.

Generation dimension This dimension captures usability degradation. We assign *Repetition* when the response repeats the prompt or itself, while Verbosity indicates a response that is unnecessarily long without necessarily repeating content. Finally, we use *Off-topic* when the response does not properly follow the instruction or includes unrelated content. We also include a *Generation error* label covering cases of empty output or a generic generation failure.

C Training details

We show in Table 8 the full hyperparameters configurations we used to train our adapters. We also analyse the effect of training data size by randomly sampling 20%, 30%, and 50% of the full training set, stratified by domain, and fine-tuning LoRA adapters with a fixed training budget (same number of steps) across all conditions (Figures 7 to 10). Across all tasks and metrics, performance increases monotonically with the amount of training data, confirming that larger vernacular training sets consistently benefit adaptation. Notably, no model appears to plateau within the range we tested, suggesting that our adapters are likely in an underfitting regime and would benefit from additional training data.

The one exception to this trend is summarization coverage, which follows a U-shaped curve.We hypothesize that this pattern is related to the fixed training budget: at lower data percentages, the model is exposed to each example more frequently, which may reinforce Gemini's compressed summary style before the model has seen enough variety to learn content selection. Coverage, unlike faithfulness, requires understanding what information is salient in a given context, a harder skill to acquire from a small, repeatedly seen set of examples. As data increases, sufficient variety is introduced for the model to begin learning this aspect of the task as well.

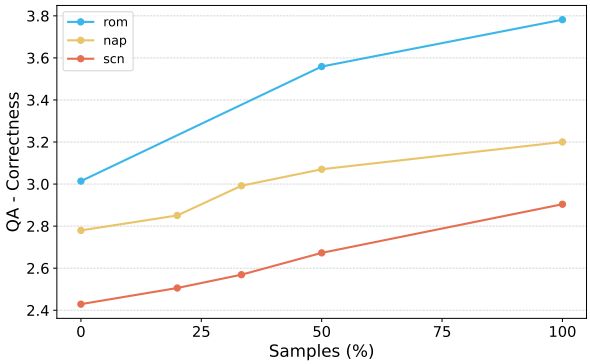

Figure 7: QA performance on different percentage of training data. Training budget remains the same for each run.

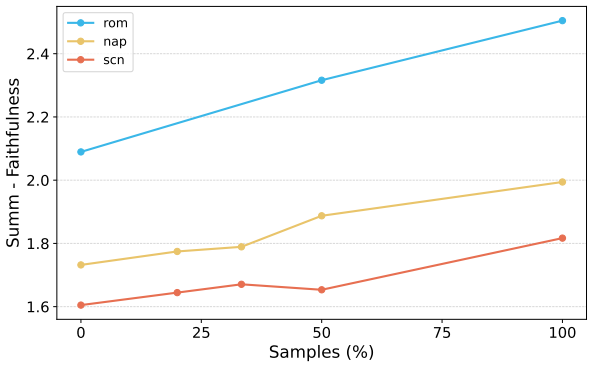

Figure 8: Summarization (Faithfulness) performance on different percentage of training data. Training budget remains the same for each run.

All experiments were carried out on machines equipped either with 2 RTX 4090 or 4 A100, for a total consumption of ∼144 GPU hours, considering both training and inference.

D Full results

Summarization and MT results Figures 4 and 11 reports summarization performance across dialect varieties and model groups. Results follow

Dimension	Label	Description
Language	Lexical leakage	Contains vernacular lexical words where a standard Italian equivalent exists.
	Grammatical leakage	Contains vernacular grammatical structures.
	Vernacular output	Response is fully or predominantly in the vernacular variety.
	Language error	Contains grammatical issues unrelated to vernacular influence.
Meaning	Hallucination	Contains invented, unsupported, or factually incorrect information.
	Translation error	Translation error that alters or significantly distorts the intended meaning.
Generation	Repetition	Prompt or answer repetition.
	Verbosity	Response is unnecessarily long without repeating content.
	Off topic	Response does not follow the instruction or includes unrelated content.
	Generation error	Empty output or generic generation error.

Table 6: Annotation scheme used to evaluate model responses before and after fine-tuning across three independent dimensions. The label set is non-exhaustive.

Variety	Area	Age range	Exposition	Speaking	Writing
ROM	Rome	50–60	5	4	2
ROM	Rome	20–30	5	4	2
ROM	Rome	20–30	4	3	2
NAP	Benevento	20-30	4	3	1
NAP	Naples	30–40	4	3	1
NAP	Naples	50-60	4	4	1
SCN	Messina	20-30	5	5	1
SCN	Catania	20-30	3	3	2
SCN	Ragusa	40-50	3	3	1

Table 7: Demographic information on the recruited native speakers. We asked them to self-asses their exposition to the target vernacular language and their ability at speaking and writing it. Notably, their self-reported writing skills are much lower compared to their speaking ones, consistently with the fact of vernacular languages being predominantly oral.

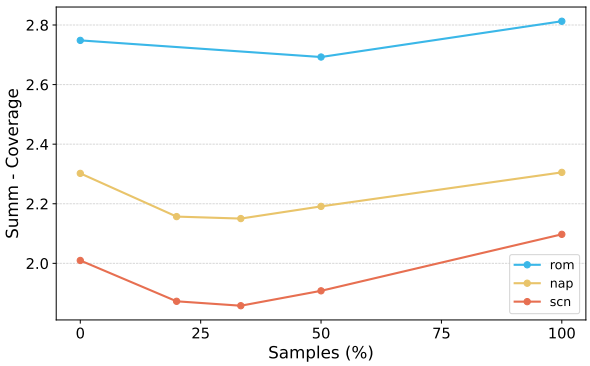

Figure 9: Summarization (Coverage) performance on different percentage of training data. Training budget remains the same for each run.

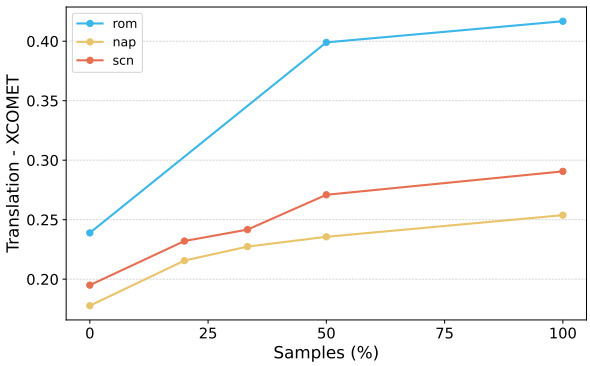

Figure 10: Translation performance on different percentage of training data. Training budget remains the same for each run.

the same ROM > NAP > SCN hierarchy observed for QA, and model group rankings are consistent with the main findings. A notable pattern concerns the coverage dimension: while all models improve on faithfulness after LoRA fine-tuning, coverage scores decrease relative to zero-shot. We attribute this to a length effect induced by the synthetic training data. As shown in Table 9, zero-shot summaries average 116 tokens, decreasing to 96 under few-shot prompting and to 44 after LoRA fine-tuning, closely matching the average length of the Gemini-generated reference summaries (41 tokens). This

Config ID	Rank	Learning Rate	Epochs
S1	8	2×10^{-5}	3
S2	8	5×10^{-5}	3
S3	8	1×10^{-4}	3
M1	16	2×10^{-5}	3
M2	16	5×10^{-5}	3
M3	16	1×10^{-4}	3
M4	16	3×10^{-4}	3
M5	16	4×10^{-4}	3
M6	16	4×10^{-4}	5
M7	16	1×10^{-4}	5
M8	16	3×10^{-4}	5
L1	32	2×10^{-5}	3
L2	32	5×10^{-5}	3
L3	32	1×10^{-4}	3
L7	32	1×10^{-4}	5
L8	32	3×10^{-4}	5

Table 8: Full list of the training configurations used in our experiments. We kept a fixed $\alpha = \text{rank} \cdot 2$.

Setting	**Avg length (words)**
Zero-shot prompting	116.12
Few-shot prompting	95.98
After LoRA finetuning	44.17
Reference summaries	41.11

Table 9: Average generated summary length under different settings. The reference summaries were generated with Gemini and used as gold references. Reported values are averages across all evaluated models.

Variety	**QA**	**Summ.**	**MT**
Minor leakage	8.33	2.67	23.00
Major leakage	0.33	–	3.00
Hallucination	1.00	1.67	0.33
Translation error	2.17	1.67	20.00
Verbose	–	0.33	–
Off-topic	0.50	0.33	0.33
Generation error	1.67	0.33	0.33

Table 10: Gemini error rates (%) by task. We group them by *language issues* (first block), *meaning issues* (second block), and *generation issues* (third block).

suggests that LoRA adapters inherit the compression style of the synthetic training targets, producing more concise and faithful summaries at the cost of coverage. The faithfulness–coverage tradeoff is therefore largely an artifact of the training data distribution rather than a limitation of the adaptation approach itself. Figures 5 and 12 reports machine translation results. As discussed in Section 5.1, translation follows the ordering ROM > SCN > NAP, with scores on the OOD control set (CODET) substantially higher due to shorter segment length.

Task difficulty QA scores are systematically higher than summarization scores across all conditions, consistent with the relative difficulty of the two tasks: QA requires identifying a correct answer span, while summarization demands both faithfulness to the source and adequate coverage of its content. MT scores on our test set are significantly lower than those obtained on the out-of-distribution CODET[9] set, which we attribute primarily to segment length (∼6 words), which reduces the translation problem to near word-level substitution, whereas our test set contains longer and syntactically complex utterances, better reflecting real-world contexts.

Effect of few-shot prompting Few-shot prompting yields consistently little benefit across tasks, and in several conditions slightly degrades performance relative to zero-shot. This is unsurprising: in-context learning can transfer task-solving strategies, but it cannot inject the parametric knowledge of a linguistic variety that the model lacks. This finding motivates our use of LoRA adaptation as the primary intervention and establishes few-shot as a useful additional comparison point.

Control settings A question-only control (QA only) and an Italian-context condition (QA and summarization) further support the variety hierarchy. In the question-only control, performance is near-minimum overall, confirming that models lack parametric knowledge of these varieties; the exception is SCN, which scores highest among the three, likely reflecting its greater presence in web-crawled pretraining data. In the Italian-context condition, SCN consistently yields the lowest scores even when the vernacular barrier is removed, corroborating annotator observations that SCN texts are internally more variable and harder to process independently of language. These results suggest that SCN's lower performance in the main experiments is driven by linguistic and data factors rather than by limited pretraining exposure.

[9] We filter sentences containing less than 3 words.

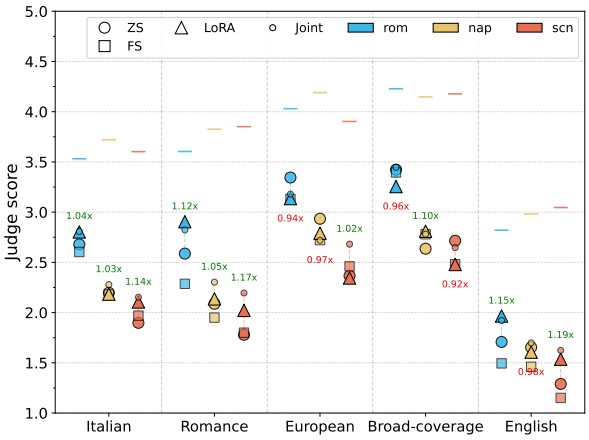

Figure 11: Summarization performance on Coverage grouped by our model classes. We show the increment of LoRA finetuned models with respect to the zero-shot baseline. Horizontal lines are the control settings Italian-context.

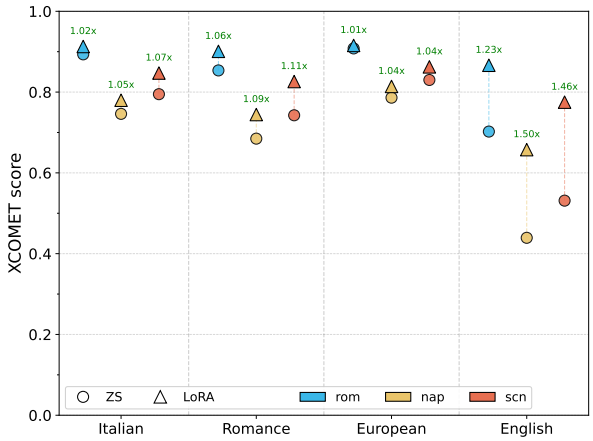

Figure 12: MT performance scored by XCOMET grouped by our model classes on the OOD dataset CODET. We show the increment of LoRA finetuned models with respect to the zero-shot baseline.

Variety	ROM	NAP	SCN
Minor leakage	8.75	9.50	13.50
Major leakage	0.75	0.75	1.25
Hallucination	0.25	1.00	1.75
Translation error	5.75	8.25	5.50
Verbose	0.25	–	–
Off-topic	0.25	0.25	0.75
Generation error	0.50	1.75	0.75

Table 11: Gemini error rates (%) by dialect. We group them by *language issues* (first block), *meaning issues* (second block), and *generation issues* (third block).

E Qualitative Error Analysis

We analyse errors in both the Gemini-generated SFT data and the outputs of our trained models, as identified and categorized by a linguist during the annotation phase (Section 3.3) following the guidelines in Table 6.

Gemini errors Gemini produces the most errors on translation, particularly for Neapolitan and Sicilian, which also exhibit lower vernacular leakage rates than Roman (Tables 10 and 11). The higher translation error rate for the two southern varieties is consistent with their greater lexical distance from Italian (Figure 1 and table 5), and extends the findings of Sprugnoli et al. (2025), who observed similar difficulties at the word level for Sicilian, to the sentence and paragraph level, and to Neapolitan.

Model errors Comparing zero-shot, LoRA, and joint LoRA outputs on a shared set of examples (Tables 12 and 13), we observe a consistent reduction in generation issues (verbosity, off-topic outputs, generation errors) and in major vernacular leakage – the most problematic error type, involving mixing of variety and Italian. Minor leakage slightly increases, which the annotating linguist attributed to names and titles that should legitimately remain in the respective variety. Hallucinations decrease across varieties except for Neapolitan, where a slight increase is observed. Roman exhibits the most errors in zero-shot, contrary to what Gemini's error profile might suggest, probably due to domain difficulty rather than linguistic factors. Summarization shows the strongest reduction in leakage across tasks, but the highest increase in hallucinations, consistent with its greater generative demands. The joint LoRA largely mirrors the behaviour of the single variety-specific adapters.

F Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Claude in order to perform the following operation: grammar and spelling check, formatting assistance, writing style improvement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

G Intended Use

The LoRA adapters, synthetic training data, and evaluation corpus produced in this work are intended solely for research purposes. The adapters were trained on synthetic data derived from dialectal source texts publicly available on the web.

Annotation	ROM			NAP			SCN		
	ZS	LoRA	J. LoRA	ZS	LoRA	J. LoRA	ZS	LoRA	J. LoRA
Minor leakage	16.50	17.50	16.50	4.57	6.29	5.14	13.00	19.00	20.00
Major leakage	8.00	3.20	3.20	8.00	4.00	5.60	9.33	4.00	5.33
Hallucination	50.00	38.00	40.00	9.33	13.33	9.33	34.00	24.00	22.00
Translation error	26.67	21.33	18.67	21.33	16.00	18.67	16.00	22.67	25.33
Verbose	8.00	0.00	0.00	16.00	2.00	0.00	28.00	0.00	0.00
Off-topic	28.00	12.00	16.00	10.00	2.00	4.00	8.00	2.00	2.00
Generation error	10.00	2.00	2.00	2.67	4.00	6.67	12.00	1.33	0.00

Table 12: Frequency (%) of the error types in the annotated subset by dialect. We group them by *generation issues* (first block), *language issues* (second block), and *meaning issues* (third block).

Annotation	QA			Sum.			Transl.		
	ZS	LoRA	J. LoRA	ZS	LoRA	J. LoRA	ZS	LoRA	J. LoRA
Minor leakage	8.67	15.33	16.00	13.60	7.20	4.00	12.00	16.50	16.50
Major leakage	7.33	4.67	5.33	6.00	4.00	8.00	10.40	2.40	2.40
Hallucination	30.67	17.33	17.33	32.00	37.33	33.33	8.00	0.00	0.00
Translation error	4.00	8.00	8.00	24.00	20.00	25.33	36.00	32.00	29.33
Verbose	20.00	1.33	0.00	8.00	0.00	0.00	–	–	–
Off-topic	25.33	4.00	5.33	5.33	6.67	9.33	–	–	–
Generation error	2.67	1.33	2.67	8.00	2.00	4.00	13.33	4.00	2.67

Table 13: Frequency (%) of the error types in the annotated subset by task. We group them by *generation issues* (first block), *language issues* (second block), and *meaning issues* (third block).

Ranking	Model	Appearances	Weighted Score
1	EuroLLM 9B	37	78
2	Gemma2 9B	46	93
3	Occiglot 7B EU5	31	130
4	ANITA 8B	33	151
5	Llama 3.1 8B LAPT	45	167
6	Salamandra 7B	35	176
7	Occiglot 7B IT-EN	40	189

Table 14: Overall model ranking across all tasks, dialects, and runs. **Appearances** counts how many times a model appears in the top 7 across all evaluated conditions (task, dialect, run). **Weighted Score** is the sum of positional scores across all top-7 appearances: a first-place finish contributes 1 point, a second-place finish contributes 2, up to 7 points for last place, so that lower scores indicate consistently higher-ranked models. The top 7 is selected by appearance frequency and sorted by weighted score (lower is better).

Task	Source text chunks (VL)	Generated task items (SL)
QA	**ROM**: *Tutto er popolo voleva pe' fforza sarvà' la Bbella Cenci; e si nun fussino stati li sordati je sarebbe ariuscito. Ma fu ttutto inutile, perchè ggiustizia fu ffatta; a quela pôra fija, pe' ggrazia speciale je fu sortanto tajata la testa. Dice che llei prima de morì', ffece sapé' ar papa che ssi je sarvava la vita, averebbe fatto rifà' ttutto de nôvo co' la chiara de ll'ôva, Ponte Rotto. Ma er papa, ora sorda!; nun intese gnente; e ddoppo fatta la carnificina se scirpò ttutti li bbeni de la famija Cènci, e cciaricchì la su' famija. Dice che fra la robba che s'aranciò cc'era puro Villa Bborghese. Anzi la maggior parte de queli bbeni agnédono a ffinì' ttutti in mano de li principi Bborghesi che èreno, nun so pe' pparte de chi, pparenti der papa.*	**ITA**: Question: Cosa disse che avrebbe fatto la protagonista se gli avessero risparmiato la vita? Answer: Avrebbe fatto rifare tutto di nuovo con il chiaro dell'uovo.
Summ.	**SCN**: *Li jatti arrivaru 'n cunzicutanza e, a leggiu a leggiu, si fìciru allatinari di l'omu puru comu armali di cumpagnìa. Pi l'Anticu Eggittu, la jatta era n'armali sacru. La dia Bastet, patruna di l'amuri, avìa la testa di jatta. S'ànnu truvatu assai mummi di jatti ntê vurbicati. Ntâ Grecia foru purtati dî Finici chi ci l'avìanu arrubbatu all'Eggizziani. Prima di tannu, li greci canuscìanu sulu furetti e baddhottuli pâ caccia ê surci. Si sapi puru d'un mircatu di jatti a Ateni. A Roma la jatta era na passiuni, prima dî ricchi e appoi di tutti li genti dâ Mperiu, d'unni si spannìu pi tuttu lu munnu. Pi l'Islam, la jatta èni n'armalu assai pusitivu, grazzi â storia dâ jatta di Mahumettu, Muezza, ca lu sarbau d'un muzzicuni di scurzuni. Nfatti è l'ùnicu armalu ca po tràsiri tranchillu ntê mischiti.*	**ITA**: Il testo narra la storia del gatto, dalla sua addomesticazione e venerazione come animale sacro nell'Antico Egitto alla sua diffusione in Grecia, a Roma e poi in tutto il mondo, assumendo un ruolo positivo anche nell'Islam.
MT	**NAP**: *Lucrine è nu rione d"o commune 'e Puzzuolo ca piglia 'o nomme sujo d"o Laco 'e Lucrine, 'o stesso laco piglia 'o nomme sujo â parola latina Lucrum e zoè- abbuscà, nifatti 'o senatore Orata 'int' 'a Romma antica 'a aveva trasfurmato pe na starza 'e piesce, essenno 'o stesso laco aunite trammeto nu cunnutto cu 'o mare. Vicine ô laco 'e Lucrine nc'è 'o laco d'Averno (d"o grieco "senz' aucielle" pecché 'o ggas zurfegno accideva ll'aucielle), tanto fammuso int'â antichetate pecché 'o se crereva 'a porta d"a casacauda, mo 'o stesso laco mmece è uno d"e poche ad essere prupietà privata pur essenno preno 'e storia ed 'e scummuagliate antecarie 'e granne nteresse. Aiza ê pere d"o vurcano Monte Nuovo. Apprimma 'e ll'eruzzione d"o Montenuovo d"o 1538 addó mo c'è Lucrine c'era Tripergole, na casale termale. Ll'eruzzione 'o sgarrupette cumpletamiente.*	**ITA**: Lucrino è un quartiere del comune di Pozzuoli che prende il suo nome dal Lago di Lucrino, lo stesso lago prende il suo nome dalla parola latina Lucrum e cioè - guadagnare, infatti il senatore Orata nell'antica Roma lo aveva trasformato in una peschiera, essendo lo stesso lago unito tramite un condotto con il mare. Vicino al lago di Lucrino c'è il lago d'Averno (dal greco "senza uccelli" perché il gas solfureo uccideva gli uccelli), tanto famoso nell'antichità perché si credeva fosse la porta dell'oltretomba, ora lo stesso lago invece è uno dei pochi ad essere proprietà privata pur essendo pieno di storia e di scoperte archeologiche di grande interesse. Si erge ai piedi del vulcano Monte Nuovo. Prima dell'eruzione del Monte Nuovo del 1538 dove ora c'è Lucrino c'era Tripergole, un borgo termale. L'eruzione lo distrusse completamente.

Table 15: Examples from the framework covering QA, summarization and translation tasks. The first example is in Roman, the second is in Sicilian and the last one is in Neapolitan

Italian guidelines	**English translation**
Data structure	
Dato un insieme di testi scritti dialetto, ognuno ha allegato: • una traduzione in italiano standard • un riassunto in italiano • due coppie domanda/risposta in italiano, riguardo il testo	Given a set of vernacular source texts, each text has associated: • a translation into standard Italian • a summary in Italian • two question/answer pairs in Italian, about the vernacular text
Verification criteria	
Verificare che: • La traduzione sia corretta e fedele al testo • Il riassunto riporti i punti fondamentali del testo senza inventare fatti • Le domande sono ben poste e rilevanti rispetto al testo • Le risposte sono rilevanti rispetto alla domanda e coerenti rispetto al testo	Verify that: • The translation is accurate and faithful • The summary covers the key points of the text without fabricating facts • The questions are well-formed and relevant to the text • The answers are relevant to the question and consistent with the text
Annotation interface	
Accanto a ogni testo in italiano è presente un check su cui scegliere Sì/No. Nel caso in cui si scelga No, bisogna correggere gli errori trovati modificando il testo in italiano corrispondente.	Next to each Italian text there is a checkbox where the annotator should select Yes/No. If No is selected, the annotator must correct the errors found by editing the corresponding Italian text.

Table 16: Annotation guidelines given to each native speaker

Italian Instruction	English Translation
Question Answering task	
Rispondi alla domanda utilizzando le informazioni fornite nel testo.	Answer the question using the information provided in the text.
In base al testo fornito, rispondi alla domanda.	Based on the provided text, answer the question.
Leggi il testo e rispondi correttamente alla domanda.	Read the text and answer the question correctly.
Usa il contenuto del testo per rispondere alla domanda.	Use the content of the text to answer the question.
Fornisci una risposta basata esclusivamente sul testo dato.	Provide an answer based exclusively on the given text.
Summarization	
Puoi riassumere questo testo?	Can you summarize this text?
Qual è il riassunto del testo fornito?	What is the summary of the provided text?
Di cosa parla il testo?	What is the text about?
Puoi fornire un breve riassunto di questo testo?	Can you provide a brief summary of this text?
Riassumi il contenuto principale del testo fornito.	Summarize the main content of the provided text.
Riassumi brevemente questo testo	Briefly summarize this text.
Puoi sintetizzare il testo?	Can you synthesize the text?
Fornisci una sintesi di questo testo dialettale.	Provide a summary of this dialectal text.
In breve, cosa dice il testo?	In short, what does the text say?
Puoi spiegare brevemente il contenuto del testo?	Can you briefly explain the content of the text?
Machine Translation task	
Traduci questo testo in italiano	Translate this text into Italian.
Traduci in italiano	Translate into Italian.
Puoi tradurre questo testo in italiano?	Can you translate this text into Italian?
Traduci dal dialetto all'italiano	Translate from dialect to Italian.
Puoi tradurre questo testo dal dialetto all'italiano?	Can you translate this text from dialect to Italian?
Qual è la traduzione in italiano di questo testo dialettale?	What is the Italian translation of this dialectal text?
Traduci in italiano questo testo scritto in dialetto.	Translate this dialect text into Italian.
Come si traduce in italiano questo testo in dialetto?	How is this dialect text translated into Italian?
Riscrivi il testo fornito dal dialetto all'italiano.	Rewrite the provided text from dialect into Italian.
Converti questo testo dal dialetto locale all'italiano.	Convert this text from the local dialect into Italian.
Puoi fornire la versione italiana di questo testo dialettale?	Can you provide the Italian version of this dialectal text?
Come sarebbe questo testo in italiano?	What would this text look like in Italian?

Table 17: Task-specific list of instructions used to have more variety. At training time, for each training item we sample an instruction and add it to the prompt

Model	EN (%)	IT (%)	Rom. (%)	#Rom.	Other (%)	#Lang.
Italian-specialized						
Minerva 7B v1.0	54	46	46	1	0	2
Minerva 7B v1.5-R1	63	30	30	1	7	2
Minerva 7B v1.5-R2	45	52	52	1	3	2
Llama 3.1 8B (SAVA/LAPT) ANITA	25	75	0	1	—	2
Romance-specialized						
Lucie 7B v1.0	26	5	53	3	20	5
European-specialized						
EuroLLM 9B	48	5	19	6	33	35
Teuken 7B	42	5	26	5	32	24
Salamandra 7B	39	2	30	8	31	35
English-specialized						
OLMo3 7B	100	0	0	0	0	1

Table 18: Pretraining data composition per model. EN (%) and IT (%) denote the fraction of English and Italian tokens; Romance (%) is the aggregate share of all Romance languages; #Rom. is the number of distinct Romance languages; Other (%) covers all remaining languages; #Lang. is the total number of distinct languages. **Bold rows** indicate backbone models subsequently adapted to Italian or Romance languages by the models above; their pretraining figures thus reflect the continual-pretraining adaptation phase rather than the original base pretraining. Entries marked '—' have undisclosed details; ≈ values are approximate.

Model	EN (%)	IT (%)	Rom. (%)	#Rom.	Other (%)	#Lang.
Italian-specialized						
Minerva 7B v1.0	44	36	36	1	21	2
Minerva 7B v1.5-R1	97	3	3	1	0	2
Minerva 7B v1.5-R1	97	3	3	1	0	2
Occiglot 7B it-en	50	50	50	1	0	2
Llama 3.1 8B LAPT	97	3	3	1	0	2
Llama 3.1 8B SAVA	97	3	3	1	0	2
ANITA 8B	69	31	31	4	0	2
Romance-specialized						
Lucie 7B v1.0	92	0	8	1	0	2
Occiglot 7B es-en	50	0	50	1	0	2
Occiglot 7B fr-en	50	0	50	1	0	2
Occiglot 7B eu5	20	20	60	3	20	5
European-specialized						
EuroLLM 9B	44	5	20	6	36	31
Salamandra 7B	67	0	32	4	1	6
Teuken 7B	60	0	0	0	40	2
English-specialized						
OLMo3 7B	100	0	0	0	0	1

Table 19: SFT data composition per model. EN (%) and IT (%) denote the share of English and Italian training samples; Rom. (%) is the combined share of all Romance languages; #Rom. is the number of distinct Romance languages present; Other (%) covers all remaining languages; #Lang. is the total number of distinct languages.

QA – Correctness Prompt

You are an expert QA evaluator. Your task is to assess whether a candidate answer correctly addresses the question, using the reference answer and the source text as ground truth.

- **Score 1:** The candidate answer is factually wrong or completely unrelated to the question.
- **Score 2:** The candidate answer attempts to address the question but contains major factual errors or omits the core information present in the reference.
- **Score 3:** The candidate answer is partially correct: it captures some key information from the reference but omits important details or includes minor inaccuracies.
- **Score 4:** The candidate answer is correct and complete but includes minor unverifiable details not present in the reference or source text.
- **Score 5:** The candidate answer is fully correct and complete: it matches the reference answer in substance, or is a faithful paraphrase that preserves its full meaning.

Figure 13: Evaluation rubric: QA – Correctness

Summarization – Faithfulness Prompt

You are an expert evaluator of text summaries. Your task is to assess whether a candidate summary is faithful to the original text, i.e. it does not introduce information that cannot be verified against the source.

- **Score 1:** The candidate summary is mostly or entirely fabricated: it contains claims that directly contradict or are completely absent from the original text.
- **Score 2:** The candidate summary contains several statements that cannot be verified against the original text, significantly distorting its content.
- **Score 3:** The candidate summary is mostly faithful but includes one or two statements that add unverifiable information or slightly distort the source.
- **Score 4:** The candidate summary is faithful to the original text with only negligible unverifiable details that do not alter the overall meaning.
- **Score 5:** The candidate summary is fully faithful: every statement can be directly verified against the original text, with no hallucinations.

Figure 14: Evaluation rubric: Summarization – Faithfulness

Summarization – Coverage Prompt

You are an expert evaluator of text summaries. Your task is to assess whether a candidate summary captures the most important information from the original text, using the reference summary as a guide to what is considered salient.

- **Score 1:** The candidate summary omits almost all key information from the original text.
- **Score 2:** The candidate summary captures only a small portion of the salient points, missing most essential content.
- **Score 3:** The candidate summary covers some key points but omits several important pieces of information present in the reference.
- **Score 4:** The candidate summary covers most key information with only minor omissions of secondary details.
- **Score 5:** The candidate summary comprehensively covers all key information from the original text, consistent with the reference summary.

Figure 15: Evaluation rubric: Summarization – Coverage