

Dromedario 3: Localizing the Tülu 3 Dataset into Italian

Luca Gíoffré^{1,*†}, Marina Iuliana Aur^{1,†}, Francesco Ortame¹, Luca Moroni¹,
Alberte Fernández-Castro^{1,2}, Elena Marafatto¹ and Roberto Navigli^{1,2}

¹*Sapienza NLP Group, Dip. di Ingegneria Informatica, Automatica e Gestionale, Sapienza University of Rome, Via Ariosto, 25 - 00185, Rome, Italy*

²*Babelscape, Via Cavour, 256 - 00184, Rome, Italy*

Abstract

Despite growing interest in supervised fine-tuning (SFT) datasets, nearly all existing resources are English-only or English-dominated, leaving Italian without a comparable resource. To bridge this gap, we introduce DROMEDARIO 3, a large-scale Italian instruction-following dataset of approximately 630K translated instruction-response pairs derived from the TüLU 3 SFT mixture. Rather than translating the entire mixture, we first apply a multi-stage filtering pipeline. Specifically, we combine source-dataset filtering with automatic task and domain classification based on the SUPER-NATURALINSTRUCTIONS taxonomy to exclude language-bound tasks, while existing non-English instances are retained untranslated to preserve multilingual diversity. The selected instances are translated using TranslateGemma-27B and post-processed to remove faulty outputs. Manual validation confirms the reliability of both the automatic classification (accuracy above 87% for task labels and 95% for domain labels) and the machine translations (correctness above 95%). We further conduct a mixing study evaluating five SFT recipes with varying Italian-to-English compositions on parallel Italian and English benchmarks using two models, Llama 3.1 8B and Minerva 7B. The model with stronger Italian pretraining benefits substantially from Italian SFT data in generative settings, while recipes combining both languages offer the most robust trade-off between gains in Italian and preservation of English capabilities.

Keywords

SFT dataset, instruction tuning, machine translation, LLMs

1. Introduction

Instruction tuning has become a standard step in the development of capable Large Language Models (LLMs), enabling them to follow diverse user requests across tasks and domains [1, 2]. The release of openly licensed instruction-following datasets (collections of prompt-response pairs designed to teach models to follow user instructions), such as UltraChat and TüLU 3 [3, 4], has made this process accessible to the research community and enabled reproducible fine-tuning experiments on strong base models. However, the vast majority of such datasets are English-only, leaving non-English languages severely underserved in the instruction-tuning landscape.

This imbalance has practical consequences: models fine-tuned exclusively on English instruction data tend to perform worse on non-English tasks even when built on multilingual base models [5, 6], with larger gaps on tasks requiring cultural or linguistic specificity [7, 8, 9]. At the same time, collecting high-quality instruction data directly in other languages is difficult and expensive: human annotation at the scale required by modern SFT datasets demands substantial resources that are rarely available outside large industrial labs.

Italian exemplifies this gap: it is a major European language with a large speaker population and has attracted growing interest from the NLP research community, as evidenced by the increasing number of dedicated models [10, 11, 12, 13, 14] and evaluation benchmarks [15, 16]. Yet, the resources available for

Twelfth Italian Conference on Computational Linguistics (CLiC-it 2026), September 14–16, 2026, Palermo, Italy

*Corresponding author.

†These authors contributed equally.

✉ gioffre@diag.uniroma1.it (L. Gíoffré); aur@diag.uniroma1.it (M. I. Aur); ortame@diag.uniroma1.it (F. Ortame); moroni@diag.uniroma1.it (L. Moroni); castro@diag.uniroma1.it (A. Fernández-Castro); marafatto@diag.uniroma1.it (E. Marafatto); navigli@diag.uniroma1.it (R. Navigli)

🆔 0009-0007-9705-6797 (L. Gíoffré); 0009-0007-8672-872X (F. Ortame); 0009-0006-1210-5098 (L. Moroni); 0000-0003-3831-9706 (R. Navigli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

instruction tuning in Italian remain limited in scale and task diversity, and often lack methodological transparency.

Filling this gap requires more than translating existing resources. A naïve translation of an English instruction dataset risks preserving *linguistically ill-posed tasks* (such as English wordplay or morphological exercises), which would propagate translation artifacts into model training [5, 17]. Moreover, using current LLMs for translation can be challenging, as they can fail in a characteristic way when applied to instruction-following data: instead of translating the task, they attempt to solve it, producing responses instead of faithful target-language renderings of the source prompt. Addressing these failure modes requires explicit pipeline design, as off-the-shelf MT capabilities alone are insufficient.

In this paper we introduce DROMEDARIO 3, a large-scale Italian instruction-following dataset derived from TüLU 3 via a principled multi-stage filtering and translation pipeline. Instead of translating the entire mixture indiscriminately, we first apply a series of filtering steps to identify the subset of data that can be reliably translated into Italian while preserving its original intent. This process filters the instances based on their source dataset, their task type, domain labels, and language to determine which items should be translated and which should be retained in their original form. We then translate the selected instances into Italian and apply additional quality-control procedures to ensure the consistency and integrity of the resulting dataset. The result is a translated dataset of approximately 630K instruction-response pairs covering a broad range of tasks and domains, suitable for supervised fine-tuning of Italian and multilingual models.¹

To provide practical guidance for practitioners on integrating Italian data into SFT pipelines, we also conduct an empirical study examining how different mixtures of Italian and English SFT data affect model performance. Specifically, we evaluate multiple language mixing ratios on ITA-Bench [15], an Italian-English benchmark suite, using two models: Llama 3.1 8B [18], an open-weight, English-first model, and Minerva 7B [10], an open-data Italian-English bilingual model.

Our contributions are as follows:

1. **A scalable methodology** for cross-lingual SFT dataset adaptation, including explicit handling of language-bound task types and domains to preserve task validity under translation.
2. **DROMEDARIO 3**, the largest publicly documented multi-source Italian SFT dataset to date, comprising $\sim 630\text{K}$ instruction-response pairs derived from TüLU 3 and annotated with task-type and domain labels.
3. **An empirical data-mixing study** quantifying the effect of Italian SFT data on both Italian- and English-language task performance across two models and multiple data recipes.

2. Related Work

2.1. Instruction Tuning and Multilingual Adaptation

The landscape of supervised fine-tuning datasets has expanded rapidly since the introduction of InstructGPT [1] and Self-Instruct [2], which established the paradigm of aligning language models to follow natural language instructions. Early datasets such as FLAN [19] relied on reformatting existing NLP benchmarks into instruction format; subsequent work moved toward more diverse, open-ended conversational data. UltraChat [3] introduced large collections of synthetically generated multi-turn dialogues, while WizardLM [20] proposed a method for automatically generating instructions of increasing complexity. OpenOrca [21] and OpenHermes [22] extended this line with distillation-based approaches using GPT-4 outputs as a supervision signal. More recently, TüLU 3 [4] provided a systematic, openly licensed collection combining multiple data sources with principled mixing, serving as the basis for DROMEDARIO 3. Despite this abundance, virtually all large-scale SFT datasets are English-only or heavily English-dominated, limiting their direct applicability to non-English language model development.

¹We release the code and dataset at https://github.com/sapienzanlp/Dromedario_3.

To address this limitation, several efforts have sought to close the multilingual gap in instruction-tuning data. The Aya model and dataset [6, 17] represent the most comprehensive attempt to date, covering over 100 languages through a combination of human annotation and large-scale machine translation via NLLB [23]. Their ablation studies offer directly relevant findings for translation-based approaches: upweighting translated instruction data significantly improves performance on open-ended generative tasks, while template-based data tends to be more beneficial for discriminative benchmarks.

Shaham et al. [5] offer a more nuanced picture of how much multilingual data is actually needed: even a small amount, as few as 40 examples, or roughly 1% of the tuning set, can unlock substantial cross-lingual capabilities, including generalization to languages not present in the SFT data at all. Their findings suggest that instruction following may be a language-agnostic skill, with multilingual exposure acting as a trigger rather than a proportional contributor. Zhu et al. [24] show that cross-lingual alignment tuning via translation can extend English instruction-following capabilities to other languages, though they note the importance of filtering linguistically incompatible content.

DROMEDARIO 3 builds on this line of work while addressing two limitations of prior translation-based approaches. Rather than translating an SFT dataset uniformly, we introduce an explicit multi-stage filtering pipeline to identify which instruction categories can be transferred across languages without altering their intent. We further complement the dataset release with controlled mixing experiments that quantify the effect of Italian SFT data on both Italian and English performance.

2.2. Instruction Tuning Efforts for Italian

Compared to the English landscape above, Italian instruction data is scarce and rarely released as a documented resource in its own right. We survey here the most widely used resources rather than an exhaustive inventory. The openly published datasets are for the most part single-corpus translations, from the Alpaca-derived Camoscio [25] and Stambecco [26] to larger community efforts such as OpenItalianData,² and each inherits the task profile of the English corpus it derives from. Multilingual collections provide Italian data that is native rather than translated: OpenAssistant Conversations [27], in its OASST-1 and OASST-2 releases, and the Aya Dataset [17] are both human-authored, yet each contributes only a small Italian portion confined to open-ended dialogue. EuroBlocks [28], whose English portion absorbs TüLU 3 wholesale, draws its Italian content from unrelated sources and reports domain and language composition separately, so it cannot be established from the release which Italian instances cover code, mathematics, or verifiable constraints.

The remainder exists only as the training mixture of a released model. Minerva [10] assembles roughly 206K Italian instances from ten sources, the largest being Bactrian-X [29], an in-house Magpie [30] translation, and a safety set derived from ALERT [31], while the subsequent continual-pretraining study [11] shifts its Italian budget upstream and pairs English TüLU 3 with a ~30K Italian supplement of native, translated, and synthetic data. DanteLLM [12] mixes translated academic corpora (Italian SQuAD, Europarl) with Camoscio, while ANITA [32] tunes on a ~100K English set merging WizardLM [20], Dolly v2 [33], and Alpaca [34] variants and adapts to Italian on filtered mC4 web text,³ so no Italian instruction-response pair enters instruction tuning at all. Maestrale [35] reports scaling to 1.7M conversations and Velvet [36] over 2M SFT examples, neither with a documented composition, and EngGPT2 [14] describes a multi-stage pipeline over Italian and English reasoning modes without releasing the mixture. In most cases, the SFT data is documented in part or not at all, and sometimes is not even released.

Across these efforts, filtering, where present, operates on instance-level quality rather than on whether an instruction type transfers across languages at all, and no Italian resource pairs its content with a matched English counterpart at the instance level at a large scale. DROMEDARIO 3 addresses this gap by providing a large-scale, principled alternative grounded in an explicit taxonomy of transferable instruction types. We further expand this discussion in Appendix A, in which we situate DROMEDARIO 3 among existing Italian and multilingual instruction resources along the axes of scale, composition, and intended use.

²<https://huggingface.co/datasets/DeepMount00/OpenItalianData>

³https://huggingface.co/datasets/gstarti/clean_mc4_it

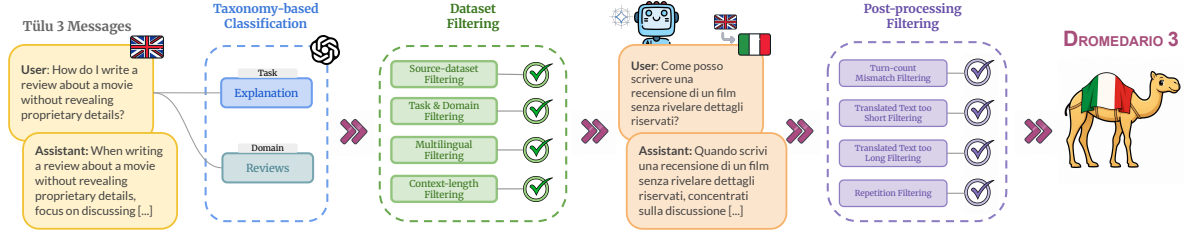


Figure 1: DROMEDARIO 3 classification, filtering and translation pipeline.

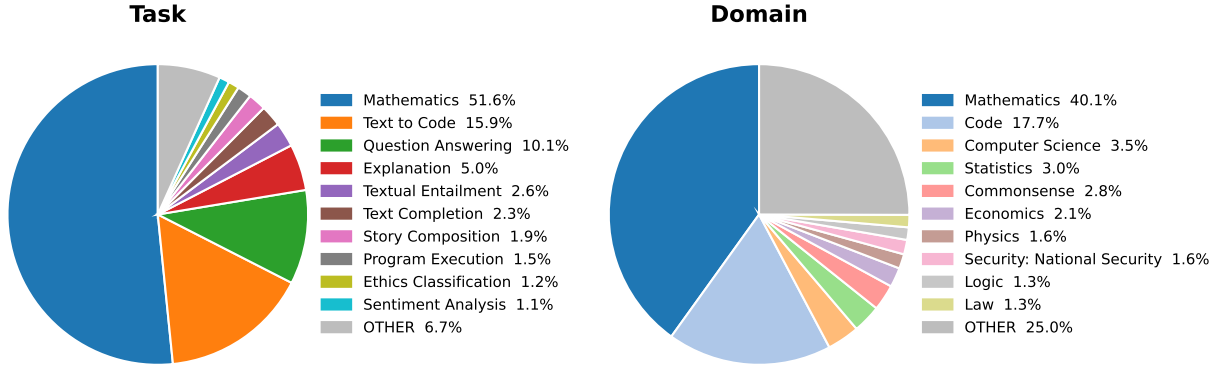


Figure 2: Distribution of task types and domains in DROMEDARIO 3. The pie charts show the top 10 most frequent task types (left) and domains (right) in the final dataset after the filtering and translation pipeline; all remaining categories are grouped under *Other*.

3. DROMEDARIO 3: Dataset Construction

In this Section, we describe the construction of DROMEDARIO 3 from the TüLU 3 SFT mixture (Figure 1). TüLU 3 is a collection of heterogeneous SFT source datasets, spanning different tasks, domains, and data-collection procedures. We first present our TüLU 3 data classification and filtering procedure, which identifies the instances suitable for translation (Section 3.1), and then describe the pipeline used to translate the selected instances into Italian (Section 3.2). Table 1 provides an overview of the filtering stages, reporting the number of removed instances at each step and the final size of DROMEDARIO 3, resulting in 627,817 out of 963,103 items. Figure 2 shows the final distribution of tasks and domains in DROMEDARIO 3 after filtering.

3.1. Data Classification and Filtering

Source-dataset Filtering Given its heterogeneous composition, we do not assume that TüLU 3 can be translated indiscriminately. Rather, we first inspect the mixture at the source-dataset level and identify datasets that are not appropriate targets for direct translation. We provide representative prompt examples for each of these decisions in Table 4 (Appendix B). This step is motivated by the observation that certain sources are inherently multilingual (e.g., to preserve cross-lingual variation) or strictly dependent on the English language for their underlying task utility. Depending on their role, we either remove these sources from the training dataset or exclude them from the translation pipeline:

- `tu_lu_hard_coded` (24K items): this dataset contains identity-related prompts designed to instill a specific model persona. Since persona-based examples are not relevant to our experimental goals, we remove the full dataset from our training set.
- `sciriff` (10K items): this dataset focuses on scientific paper comprehension tasks, which are largely tied to English-language conventions. Given that Italian scientific production is limited in volume and stylistically distinct, translated instances would risk being ill-posed. We therefore retain the entire dataset in its original form.

- **aya** (100K items): this dataset contains instruction-response pairs spanning 65 languages. We retain it in its original multilingual form, preserving its cross-lingual diversity by design. English-language Aya instances are also included to maintain task variation.
- **wildchat** (100K items): this dataset focuses on real conversations between users and chatbots, collected from public conversation-sharing platforms. Because its multilingual and highly diverse user-generated content makes systematic translation potentially unreliable, we retain the dataset in its original form, preserving its conversational diversity.
- **tablegpt** (5K items): this dataset covers a broad range of table-related tasks (e.g., table understanding, table manipulation). We retain the full original dataset to preserve its proper formatting.

In total, this step removes all 24K persona-based instances of `tu1u_hard_coded` from the training set and excludes 215K instances from the translation pipeline.

Taxonomy-based Classification After source-dataset filtering, we classify the remaining Tülu 3 instances along two distinct dimensions: task type and domain. This is motivated by the observation that not all instruction-following instances are equally suited for translation. For example, tasks and domains relying on English-specific linguistic properties can become misleading or ill-posed when translated into another language. We report the full list of task types and domains in Table 6.

In particular, we follow the SUPER-NATURALINSTRUCTIONS (SNI) taxonomy [37], one of the most comprehensive categorizations of instruction-following data, spanning over 1,600 labels grouped into 76 task types and 78 domains. The SNI classification was constructed through a large-scale community effort, collecting tasks from diverse NLP benchmarks and datasets, each accompanied by natural language instructions, positive and negative examples, and metadata including task type and domain labels. Task types capture the *nature* of the required output (e.g., Classification, Generation, Translation), while domain labels reflect the *subject* or *genre* of the input text (e.g., Medicine, Law, History).

For each instance, we use **GPT-oss-120b** [38] as our classifier; given the user prompt and the label set for the relevant dimension (task type or domain), the model assigns the most appropriate label to the instance. We treat classifications as invalid when the model output i) is not parsable, ii) contains hallucinated labels, or iii) does not assign a label. Upon manual inspection, we found that the latter case mainly affects safety-related instances, as for these items the model refuses to provide a response. Overall, invalid classifications account for 19,835 instances, corresponding to an error rate of 2%: we therefore discard these instances from the downstream label-based analyses. We provide further details and classification prompts in Appendix C.

Task & Domain Filtering Based on this classification, we perform a filtering step using the task and domain labels assigned to each sample, asking two expert linguists to annotate the *translatability* of these labels.⁴ A task or domain is classified as *translation-safe* if its intent and validity are preserved under translation; conversely, we mark a task or domain as *non-translatable* when its formulation is inherently language-specific or would become trivial or misleading when translated into Italian. Five main rationales guide the exclusion of task and domain labels from the translation pipeline:

1. *Language-form dependence*: tasks that are inherently untranslatable, as they target linguistic phenomena tied to the surface form of the source language;
2. *Translation-sensitive labels or meanings*: tasks whose labels, meanings, or target phenomena may be lost, distorted, or artificially introduced through translation;
3. *Structure or format sensitivity*: tasks that make heavy use of special tokens, formatting artifacts, structured inputs, or non-translatable elements;
4. *Methodological invalidity*: tasks for which translation would make the evaluation circular or collapse the task definition;
5. *Low-value or high-risk translation*: tasks for which translation is unnecessary, unreliable, or of limited cross-lingual value.

⁴The annotators have access to a sample for each task and domain label.

Pipeline Step	Removed	Untranslated	Data Instances
TÜLU 3			963,103
<i>Pre-processing filtering</i>			
Source-dataset Filtering	−24,000	−215,000	724,103
Unlabelled/Misclassified Instances		−19,835	704,268
Task & Domain Filtering		−45,557	658,711
Multilingual Filtering		−7,298	651,413
Context-length Filtering		−12,679	638,734
<i>Post-processing filtering</i>			
Turn-count Mismatch Filtering		−4,217	634,517
Translated Text too Short Filtering		−3,841	630,676
Translated Text too Long Filtering		−1,576	629,100
Repetition Filtering		−1,283	627,817
DROMEDARIO 3	−24,000	−311,286	627,817

Table 1

Number of instances removed from the training set (**Removed**) and filtered from the translation pipeline (**Untranslated**) after each processing step, along with the remaining number of data instances (**Data Instances**). The final row reports the total number of removed and filtered instances and the resulting size of DROMEDARIO 3.

Tables 6 and 7 in Appendix D summarize the resulting division between translatable and non-translatable labels for both tasks and domains, and provide further explanations for exclusion. Overall, we mark 40 out of 76 task types and 2 out of 78 domains as non-translatable, for a total of 45,557 ($\sim 5\%$) items.

Multilingual Filtering We exclude non-English instances from the translation pipeline while retaining them in the final mixture, leveraging FastText’s pretrained language identification model `lid.176.ftz` [39]. Since these instances already provide multilingual diversity, translating them into Italian would erase their original role of adding multilingual diversity. This step excludes 7,298 items ($\sim 0.75\%$ of the original mixture).

3.2. Translation Pipeline

Translator Selection To obtain the most faithful and fluent translations, we first conduct a preliminary investigation to select the best translator among three open-weight models from different families and scales: TowerPlus-72B [40], TranslateGemma-27B [41] and GPT-oss-120b [38]. We select the first two because they are specifically fine-tuned (TowerPlus) or pretrained (TranslateGemma) for machine translation; we additionally include GPT-oss-120b as a larger general-purpose LLM baseline.

We employ sampling stratified by source dataset to obtain a subset of 1,000 instances, translate each instance with all three candidate models, and evaluate the resulting outputs using xCOMET-XL [42], a reference-free translation quality-estimation metric. For each instance, we score the three candidate translations independently and select as the preferred output the translation receiving the highest xCOMET-XL score. We then count how often each model is preferred across the entire sample. TranslateGemma-27B achieves the highest preference rate, at $\sim 39\%$, compared with $\sim 32\%$ for GPT-oss-120b and $\sim 29\%$ for TowerPlus-72B.

To complement this quantitative evaluation, two of the authors manually inspect a sample of translations generated by the three models. We find that TranslateGemma-27B tends to produce the most faithful and fluent translations among the three, despite being the smallest model, which we tentatively attribute to its translation-specific training and to the strong multilingual foundation of the Gemma 3 family. We also observe a failure mode that is particularly problematic for instruction-following data: rather than translating the source instance, the model occasionally attempts to *solve* the underlying task, producing an Italian response to the original English prompt instead of a faithful target-language

rendering. TranslateGemma-27B is not immune to this behavior, but we observe it comparatively less often than its competitors.

These results, along with **TranslateGemma-27B**'s smaller size and lower inference cost, lead us to select it as the translation model for our pipeline.⁵

Context-Length Filtering After selecting the translator, we exclude from the translation pipeline all samples whose tokenized length exceeds 2,048 tokens, corresponding to the input context window of TranslateGemma-27B. This step excludes 12,679 instances and ensures that all remaining instances can effectively be translated without truncation.

Translation Approach We translate each instance in a single call by feeding the translator with all conversation turns, including both user prompts and assistant responses. This approach makes the translation tractable, as the model does not lose any important context during translation. However, this raises two challenges. First, we need a way to reconstruct individual turns from the model's output, since the translator receives the full conversation as a single input. Second, as noted above, when translating instruction-following data, models tend to *solve tasks* rather than translate them. We cannot address either challenge through prompt engineering, as the model only accepts a source and target language code in the user prompt, and does not support a system prompt.

To address both points simultaneously, we format each conversation as an **explicit dialogue between two users**. We prepend distinct prefixes to each turn, "User1 asks:" for user prompts and "User2 replies:" for assistant responses, serving two purposes: they act as reconstruction anchors for recovering individual turns, and they frame the input as a dialogue to be translated rather than a task to be executed. To further steer the model toward translation, we also prime the output by prefilling the generation with the Italian translation of the opening prefix (e.g., "L'utente1 chiede: "). The original turns are then recovered by splitting the output along these predefined separators, accounting for possible translation variants (e.g., "L'utente2 risponde:", "L'utente2 replica: "). Appendix E reports the full list of separator patterns and the translation prompt in Table 9 and Table 8, respectively.

Post-Processing Steps The formatting pipeline described above enables the translation of the instruction-following dataset by substantially reducing task-execution errors. After reconstructing the conversation turns, we apply a set of structural and length-based filters to exclude faulty translations.

First, we discard instances whose reconstructed turn count does not match the original dialogue structure, as this indicates either that the model solved the task or that parsing failed. Second, we remove translations with large length discrepancies relative to the source, retaining only instances whose target-to-source word-count ratio falls between 0.6 and 1.5 (found after iterative manual inspection). Very short outputs typically indicate omissions, incomplete translations, or task execution instead of translation, while overly long outputs often result from repetition or uncontrolled generation.

Finally, we exclude empty translations and generations exhibiting degenerate repetition loops with simple overlap-matching checks. In total, this step excludes 10,917 instances.

4. Dataset Manual Validation

We manually validate both the automatic classification and the translated instances. Two expert linguists fluent in both Italian and English are each given a stratified sample of 300 instances (100 instances in common) and are asked to independently annotate task and domain label, and machine translation of each item. This results in 500 unique annotated items; annotation guidelines are provided in Appendix F.

⁵We used <https://huggingface.co/Infomaniak-AI/vllm-translategemma-27b-it>, a vLLM-compatible port of the model.

4.1. Classification Validation

The annotators judge GPT-oss-120b’s labels as correct in the large majority of cases, with domain labels ($\sim 95\%$) more reliable than task labels ($\sim 87\%$). Inter-annotator agreement (IAA) is substantial for both dimensions, with Cohen’s κ values of 0.80 for task-type labels and 0.75 for domain labels.

A qualitative inspection shows that the misclassifications are systematic rather than arbitrary, since the classifier tends to assign a broader task label where a more specific one applies. The two task labels most frequently corrected by annotators are i) *Answerability Classification* (i.e., whether a question can be answered at all), often misclassified as *Question Answering* or *Data to Text*, and ii) *Ethics Classification* (i.e., whether to refuse or reason about harmful content), often misclassified as *Dialogue Generation* or *Text Completion*. A smaller cluster involves *Mathematics* and *Text to Code* instances absorbed into generic categories. Domain errors are markedly rarer, typically over-specific security labels or mismatched genres.

4.2. Translation Validation

Annotators additionally judge the correctness of each translation produced by TranslateGemma-27B and assign an error type when the translations do not follow the guidelines in Appendix F. Translations are deemed correct in $\sim 95\%$ of cases, with high IAA (Cohen’s κ of 0.85), confirming that the Italian instances generally preserve the meaning and task requirements of the English sources.

The failures cluster into a few recurring patterns (i.e., incorrect translation, incomplete translation, task execution, or other), which we describe below within these categories.

The most frequent kind of *incorrect translation* is the over-translation of content that must stay in the source language for the task to remain valid. This error affects code- and structure-oriented sources, where SQL column names or Python identifiers are translated into Italian and the resulting code no longer runs. When the translated content carries no functional role, such as a placeholder string value, localizing it into Italian does not affect the task and can be considered correct behavior. A subtler case arises when translation alters the instruction itself, for instance a prompt that required an English output but, once translated, demands an Italian one, or semantic calques flagged by the annotators (e.g., *recommended* translated as *raccomandata* instead of *consigliata*).

Less frequent are *incomplete translations*, truncated mid-sequence (most visible in multi-step Mathematics chains), and partial translations, in which only part of the instance is rendered into Italian while the rest is left in English. These error types are consistent with the phenomena addressed by the structural and length-based filters described in Section 3.2. Their limited presence in the manually inspected sample suggests that the filtering step reduces but does not eliminate the main failure modes of LLM-based translation. Table 10 in Appendix F reports representative examples for each error type.

5. Experimental Setup

SFT Recipes To evaluate the quality of DROMEDARIO 3, we design five SFT recipes to systematically compare it with the TüLU 3 dataset. Our goal is to assess whether incorporating silver-quality Italian SFT data improves model performance on Italian benchmarks and to what extent different mixing ratios affect downstream results. We use the following configurations:

1. **Control** (939,103 items): we fine-tune models on the TüLU 3 dataset after removing the 24K persona-based instances. We use this recipe as a baseline and compare our results against it.
2. **Sub-50** (939,103 items): we replace 50% of the TüLU 3 training instances that are translation-safe with samples from DROMEDARIO 3. Substitution is performed uniformly across all translated data sources to preserve the original distribution.
3. **Sub-100** (939,103 items): we replace 100% of the TüLU 3 instances that have a corresponding counterpart in DROMEDARIO 3, effectively substituting the overlapping portion of the dataset. We leave non-overlapping TüLU 3 data unchanged.

4. **Add-50** (1,253,012 items): we augment the filtered TüLU 3 dataset by adding 50% of the DROMEDARIO 3 dataset. Addition is performed uniformly across all translated data sources to preserve the original distribution.
5. **Add-100** (1,566,920 items): we fine-tune on the union of the filtered TüLU 3 and complete DROMEDARIO 3 datasets, representing the upper-bound setting for multilingual instruction tuning in our experimental design.

Models We apply our fine-tuning strategies to two foundational language models. We use the base **Llama-3.1-8B** [18], a strong multilingual open-weight model trained on 15 trillion tokens, of which 8% is non-English. While it is not specialized in Italian, we employ this model as a baseline, since Lambert et al. [4] used it to optimize the TüLU 3 mix. To assess how additional Italian instruction data interacts with a model largely pretrained on Italian data, we fine-tune the base **Minerva-7B** [10], the only open-data model pretrained on 2.5 trillion tokens balanced across Italian and English.

Training Configuration We use the LLAMA-FACTORY framework [43] to fine-tune both models on all recipes. Training is performed on 4 HPC nodes each equipped with 4 A100 GPUs with 64GB of VRAM; we achieve multi-node, multi-GPU communication using the DeepSpeed library [44]. To ensure comparability with prior work, we adopt the same training configuration described in Lambert et al. [4]: we slightly increase the total effective batch size to optimize computation. We report the full parameter list in Table 11 in Appendix G.

Evaluation Benchmarks To evaluate the fine-tuned models, we use **ITA-Bench** [15], a comprehensive benchmark suite for Italian LLM evaluation. ITA-Bench provides Italian translations of several widely adopted English benchmarks, which allows controlled multilingual comparisons.

We first consider six multiple-choice benchmarks evaluated using a perplexity-based scoring approach. For each candidate answer, we compute its conditional probability given the input prompt and select the option with the highest likelihood. The model’s final score on a benchmark is the proportion of correctly predicted answers. Specifically, we employ the following multiple-choice benchmarks: **HellaSwag** (HS) [45], which measures commonsense reasoning and sentence completion; **PIQA** [46], which assesses physical commonsense reasoning; **SciQ** [47], which evaluates scientific knowledge; **ARC-Easy** and **ARC-Challenge** [48], which contain grade-school science questions of varying difficulty; and **MMLU** [49], which probes knowledge and reasoning across several domains.

We also evaluate models on the Italian translations of two open-ended generation benchmarks available through ITA-Bench: **IFEval** [50], which measures instruction-following abilities, and **MATH** [51], a benchmark composed of competition-level mathematics problems. We adopt the *hard* version of the problems, which are classified as “Level 5” in the benchmark. For these tasks, we adopt the evaluation procedures implemented in ITA-Bench, which rely on benchmark-specific answer extraction and pattern-matching strategies to automatically score generated responses. Specifically, for IFEval, we report **instance-level strict accuracy** [50], defined as the proportion of verifiable instructions that are correctly followed. For MATH, we report accuracy, measuring the fraction of problems for which the model correctly produces the final numerical answer. We follow Moroni et al. [15] for the evaluation metrics and number of shots. We report results separately for perplexity-based and open-ended evaluations, as these settings capture different aspects of model behavior and may exhibit substantially different performance trends [52]. This distinction is particularly relevant here, since the composition of the instruction-tuning mixture has been found to affect generative and discriminative benchmarks differently [6], a divergence also observed for Italian [53].

As an additional open-ended benchmark, we use the Italian subset of **MultiLoKo** [8], a multilingual knowledge QA dataset comprising 250 question-answer pairs per language. We focus on the questions sourced natively in Italian, which target locally relevant knowledge: we also use the available human-authored English translation to enable a fully parallel comparison. MultiLoKo is scored with exact match against a list of reference answers. Model performance is evaluated using substring exact match,

Model	Recipe	Multi-choice						Open-ended		
		HS <i>0 shot</i>	PIQA <i>0 shot</i>	SCIQ <i>0 shot</i>	ARC-C <i>5 shots</i>	ARC-E <i>5 shots</i>	MMLU <i>5 shots</i>	IFEval <i>0 shot</i>	MultiLoKo <i>0 shot</i>	MATH <i>4 shots</i>
Italian										
Llama 3.1	Control	59.7	69.0	84.9	46.2	62.5	58.3	65.3	16.4	7.6
	Sub-50	58.8	68.8	83.6	46.6	61.8	59.0	64.9	14.0	4.7
	Sub-100	59.6	69.8	84.3	46.6	62.3	59.2	64.2	12.8	5.6
	Add-50	59.2	68.9	85.4	46.6	62.9	58.9	65.3	13.6	5.9
	Add-100	58.2	70.2	85.5	45.8	56.4	59.1	65.7	12.8	5.8
Minerva	Control	63.1	68.4	87.1	37.7	64.5	40.0	35.4	26.8	1.8
	Sub-50	63.2	68.8	88.5	41.0	68.8	39.6	43.5	26.8	2.1
	Sub-100	62.9	69.1	89.6	41.5	68.2	39.8	43.5	26.8	2.5
	Add-50	63.0	68.8	88.7	42.6	68.4	40.6	44.0	23.6	2.1
	Add-100	62.8	69.2	89.1	41.6	68.5	40.1	46.5	24.0	1.7
English										
Llama 3.1	Control	75.6	80.9	87.4	50.0	68.0	65.8	75.2	9.2	6.2
	Sub-50	75.4	81.1	88.0	55.0	76.7	65.1	76.3	11.2	6.1
	Sub-100	75.3	81.2	91.5	54.9	78.9	65.1	75.2	8.4	4.9
	Add-50	75.7	81.4	91.1	53.2	75.0	65.2	77.1	10.0	6.1
	Add-100	74.9	81.6	85.5	52.3	71.0	65.2	75.3	11.6	5.0
Minerva	Control	72.0	74.6	94.3	44.7	71.4	42.9	51.0	16.4	2.2
	Sub-50	71.9	74.7	94.4	46.0	74.0	42.8	53.4	15.6	2.9
	Sub-100	72.1	75.5	95.0	47.7	73.7	42.2	47.5	19.2	1.4
	Add-50	71.7	75.4	94.0	46.1	74.3	42.7	49.3	16.0	2.5
	Add-100	71.8	75.7	93.6	46.0	74.3	42.9	53.8	17.2	2.5

Table 2

Evaluation results on Italian (top) and English (bottom) benchmarks across model families and SFT recipes. Bold values indicate the best score within each model family, benchmark, and language.

where a prediction is considered correct if the gold answer appears anywhere within the generated output, relaxing the verbatim matching adopted in MultiLoKo to avoid penalizing correct answers embedded in longer responses [54, 55].

6. Results

We report absolute scores in Table 2 for Italian and English, and relative changes with respect to *Control* in Figure 3. The two models respond differently to the recipes; we analyze these differences below⁶.

6.1. Results on Italian Benchmarks

On Italian multiple-choice benchmarks (Figure 3, top row), Minerva improves once Italian data enters the mixture: ARC-C, ARC-E and SCIQ rise by 4.9, 4.3 and 2.5 points over *Control* (always different recipes), while HellaSwag and MMLU stay within 0.6 points under every recipe. The best substitution and the best addition recipe never differ by more than 1.1 points. Llama stays within 1.5 points of *Control* everywhere except ARC-E under *Add-100*.

The two models diverge on the open-ended benchmarks. Minerva improves on IFEval under every recipe (+8.1 to +11.1 points), with the addition recipes ahead by up to 3.0. By contrast, Llama stays within 1.1 points of its *Control* score, which is already substantially higher than Minerva’s (65.3 vs. 35.4). MultiLoKo moves the other way: Minerva holds at 26.8 under both substitution recipes and

⁶We report but do not analyze MATH scores, as their small values inflate the relative difference.



Figure 3: Performance differences of the four SFT recipes relative to the *Control* baseline (set to 0). We report results on Italian and English benchmarks in the top and bottom row, respectively. The vertical line separates perplexity-based evaluation (PPL) and open-ended generation (GEN).

drops to 23.6 and 24.0 under the addition ones, while Llama loses 2.4 to 3.6 points under every recipe, substitution included.

6.2. Results on English Benchmarks

On English multiple-choice benchmarks (Figure 3, bottom row), Minerva improves on ARC-C and ARC-E by 3.0 and 2.9 points over *Control*, while HellaSwag, PIQA, SCIQ and MMLU stay within 1.1 points. Llama gains more, as ARC-E rises by 10.9 points (68.0 $\rightarrow$ 78.9 under *Sub-100*), ARC-C by 5.0 and SCIQ by 4.1, while HellaSwag, PIQA and MMLU stay within 0.7 points. The substitution recipes drive most of Llama’s gains, and *Add-100* is the weakest of the four, ending 7.9 points below *Sub-100* on ARC-E and 6.0 below on SCIQ.

On the open-ended benchmarks, Minerva’s IFEval gains 2.4 and 2.8 points under *Sub-50* and *Add-100* but loses 3.5 under *Sub-100*, and its MultiLoKo varies between -0.8 and $+2.8$. Llama stays within 1.9 points of *Control* on IFEval and between -0.8 and $+2.4$ on MultiLoKo.

6.3. Discussion

Whether to introduce Italian instruction data depends on the pretraining distribution of the target model. Minerva generally benefits from the introduction of Italian data, most visibly on IFEval, the benchmark our mixtures target most directly, where it gains 8.1 to 11.1 points. Llama remains moderately stable (except for the -6.1 points on ARC-E under *Add-100*), suggesting that its English instruction tuning already transfers to Italian, leaving little headroom for Italian data to fill.

The recipes also change the standing of the two models on Italian. Under matched recipes, Minerva’s lead over Llama on ARC-E widens from 2.0 points under *Control* to at least 5.5 and on SCIQ from 2.2 to at least 3.3, while Llama’s 8.5-point lead on ARC-C narrows to at most 5.6.

Both models also improve on English multiple-choice benchmarks, and the substitution recipes produce Llama’s largest English gains despite reducing the amount of English data in the mixture. This

is complementary to Üstün et al. [6], who find translated instruction data most beneficial for generative tasks: we observe that effect in Italian, on IFEval, while the transfer back to English appears on the discriminative benchmarks instead.

For Minerva, the amount of Italian data and the way it is introduced matter little: the best 50% and 100% recipes stay within 1.1 points of each other on every Italian multiple-choice benchmark. This is consistent with reports that a small fraction of multilingual data suffices to elicit cross-lingual instruction following [5], and it means the additional translation and training cost of the 100% recipes buys little. The two modes separate only on the open-ended benchmarks, and in opposite directions, with addition ahead by 3.0 points on IFEval and substitution ahead by 2.8 on MultiLoKo, so the choice depends on which capability (roughly, instruction following or factual knowledge) is targeted.

The results are consistent with the view that alignment primarily teaches format adherence rather than content [56] and with reports that some mixtures degrade factual performance relative to the base model [57]. The ARC and SCIQ gains fit the same reading: each model’s largest gains fall in the language it was predominantly pretrained on, suggesting better access to what it already knows.

A final observation concerns the dataset rather than the recipes. Under *Sub-100*, which replaces roughly two thirds of the Tülu 3 mixture with translated data, 29 of the 32 model, language and benchmark combinations we analyze stay within one point of *Control* or improve; the exceptions are IFEval and MultiLoKo in Italian for Llama (−1.1 and −3.6) and IFEval in English for Minerva (−3.5). Replacing the majority of an instruction mixture with translated data therefore carries no measurable cost in most settings, including in the source language. This concerns downstream utility rather than translation fidelity, and our design does not isolate which sources or taxonomy categories are worth translating, since the recipes vary the amount of translated data and not its composition.

7. Conclusions

We introduce DROMEDARIO 3, a large-scale Italian instruction-following dataset derived from Tülu 3. We construct it through a multi-stage pipeline combining source-dataset filtering, task classification based on the SUPER-NATURALINSTRUCTIONS taxonomy, multilingual filtering, and post-processing to identify instances that can be reliably adapted to Italian. The resulting dataset comprises approximately 630K translated instruction-response pairs, the largest *documented* multi-source Italian SFT dataset released to date. Manual validation confirms the reliability of the automatic classification ($\sim 87\%$ for task labels and $\sim 95\%$ for domains). Annotators deem $\sim 95\%$ of the generated translations correct. The remaining errors follow systematic and interpretable patterns, suggesting that the pipeline effectively filters out problematic instances while preserving the majority of the data.

Our mixing experiments show that the value of translated SFT data depends on the base model. Minerva, pretrained on a balanced Italian-English distribution, gains up to 11.1 points on Italian instruction following, while Llama 3.1 improves at most by 1.2 points on any Italian benchmark.

At a fixed mixture size, substituting translated data for English appears to be a more economical and generally safer choice: in our experiments, it yields the largest English gains for Llama while preserving Minerva’s factual recall. By contrast, the main observed advantage of the addition recipes concerns Minerva’s Italian instruction-following performance.

We release DROMEDARIO 3 to support reproducible Italian LLM development and to facilitate future research on multilingual SFT dataset construction.

Acknowledgments

Luca Gioffré, Marina Iuliana Aur, Francesco Ortame, Alberte Fernández-Castro, Elena Marafatto, Luca Moroni, and Roberto Navigli gratefully acknowledge the support of the AI Factory IT4LIA project. Elena Marafatto acknowledges the support of the Agenzia per la cybersicurezza nazionale (ACN), which funds her National PhD in Artificial Intelligence under the programme for promotion of XLI cycle PhD research in cybersecurity. The views expressed are those of the authors and do not represent

the funding institution. We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Finally, the authors would also like to thank Leonardo Colosi for his technical support with HPC usage.

Declaration on Generative AI

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

References

- [1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), *Advances in Neural Information Processing Systems*, volume 35, Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- [2] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13484–13508. URL: <https://aclanthology.org/2023.acl-long.754/>. doi:10.18653/v1/2023.acl-long.754.
- [3] N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, B. Zhou, Enhancing chat language models by scaling high-quality instructional conversations, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, 2023, pp. 3029–3051. URL: <https://aclanthology.org/2023.emnlp-main.183/>. doi:10.18653/v1/2023.emnlp-main.183.
- [4] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, X. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, H. Hajishirzi, Tulu 3: Pushing frontiers in open language model post-training, in: *Second Conference on Language Modeling*, 2025. URL: <https://openreview.net/forum?id=i1uGbfHHpH>.
- [5] U. Shaham, J. Herzig, R. Aharoni, I. Szpektor, R. Tsarfaty, M. Eyal, Multilingual instruction tuning with just a pinch of multilinguality, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2304–2317. URL: <https://aclanthology.org/2024.findings-acl.136/>. doi:10.18653/v1/2024.findings-acl.136.
- [6] A. Üstün, V. Aryabumi, Z. Yong, W.-Y. Ko, D. D'souza, G. Onilude, N. Bhandari, S. Singh, H.-L. Ooi, A. Kayid, F. Vargus, P. Blunsom, S. Longpre, N. Muennighoff, M. Fadaee, J. Kreutzer, S. Hooker, Aya model: An instruction finetuned open-access multilingual language model, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15894–15939. URL: <https://aclanthology.org/2024.acl-long.845/>. doi:10.18653/v1/2024.acl-long.845.
- [7] A. Rao, A. Yerukola, V. Shah, K. Reinecke, M. Sap, NormAd: A framework for measuring the cultural adaptability of large language models, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, As-

- sociation for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 2373–2403. URL: <https://aclanthology.org/2025.naacl-long.120/>. doi:10.18653/v1/2025.naacl-long.120.
- [8] D. Hupkes, N. Bogoychev, MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages, CoRR abs/2504.10356 (2025). URL: <https://doi.org/10.48550/arXiv.2504.10356>. doi:10.48550/ARXIV.2504.10356.
 - [9] L. Xu, L. Moroni, R. Navigli, Cultural and knowledge biases in llms through the lens of entity-aware machine translation, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 8794–8812. doi:10.63317/3jxgnspt4srr.
 - [10] R. Orlando, L. Moroni, P.-L. Huguët Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: <https://aclanthology.org/2024.clicit-1.77/>.
 - [11] L. Moroni, T. Bonomo, L. Gioffré, L. Xu, D. Fedele, L. Colosi, A. S. Bejgu, A. Scirè, R. Navigli, What we learned from continually training minerva: A case study on Italian, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 760–774. URL: <https://aclanthology.org/2025.clicit-1.72/>.
 - [12] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let’s push Italian LLM research forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: <https://aclanthology.org/2024.lrec-main.388/>.
 - [13] C. J. Scozzaro, M. Rinaldi, G. Mittone, M. A. Stranisci, Kenji-Endo: a BabyLM, in: F. Cutugno, A. Miaschi, A. Palmero Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci (Eds.), EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, number 4195 in CEUR Workshop Proceedings, Bari, Italy, 2026. URL: <https://ceur-ws.org/Vol-4195/36.pdf>.
 - [14] G. Ciarfaglia, A. Rosanova, S. Cipolla, J. Bartoli, A. D. Domenico, C. Fioroni, A. Fontana, M. R. Scoleri, M. I. Mone, D. Franchi, M. C. D. Gaudio, A. Leodori, F. Cinti, M. Capozzi, C. Baston, F. Picariello, M. Gabusi, S. Bonura, V. Morreale, I. Bailo, EngGPT2: Sovereign, efficient and open intelligence, CoRR abs/2603.16430 (2026). URL: <https://doi.org/10.48550/arXiv.2603.16430>. doi:10.48550/ARXIV.2603.16430.
 - [15] L. Moroni, S. Conia, F. Martelli, R. Navigli, Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599. URL: <https://aclanthology.org/2024.clicit-1.67/>.
 - [16] G. Attanasio, M. La Quatra, A. Santilli, B. Savoldi, ItaEval: A CALAMITA challenge, in: F. Dell’Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 1064–1073. URL: <https://aclanthology.org/2024.clicit-1.117/>.
 - [17] S. Singh, F. Vargus, D. D’souza, B. F. Karlsson, A. Mahendiran, W.-Y. Ko, H. Shandilya, J. Patel, D. Mataciunas, L. O’Mahony, M. Zhang, R. Hettiarachchi, J. Wilson, M. Machado, L. Moura, D. Krzemiński, H. Fadaei, I. Ergun, I. Okoh, A. Alaagib, O. Mudannayake, Z. Alyafeai, V. Chien, S. Ruder, S. Guthikonda, E. Alghamdi, S. Gehrmann, N. Muennighoff, M. Bartolo, J. Kreutzer, A. Üstün, M. Fadaee, S. Hooker, Aya dataset: An open-access collection for multilingual instruction tuning, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11521–11567. URL: <https://aclanthology.org/2024.acl-long.620/>. doi:10.18653/v1/2024.acl-long.620.
 - [18] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur,

A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Young, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsim-poukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhen-de, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baeviski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feicht-enhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspe-gren, H. Goldman, H. Zhan, I. Dam-laj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A. L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan,

- R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The llama 3 herd of models, 2024. URL: <https://arxiv.org/abs/2407.21783>.
- [19] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=gEZrGCozdqR>.
- [20] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, Q. Lin, D. Jiang, WizardLM: Empowering large pre-trained language models to follow complex instructions, in: The Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=CfXh93NDgH>.
- [21] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of GPT-4, CoRR abs/2306.02707 (2023). URL: <https://doi.org/10.48550/arXiv.2306.02707>. doi:10.48550/ARXIV.2306.02707.
- [22] Teknium, OpenHermes 2.5: An open dataset of synthetic data for generalist LLM assistants, 2023. URL: <https://huggingface.co/datasets/teknium/OpenHermes-2.5>.
- [23] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, CoRR abs/2207.04672 (2022). URL: <https://doi.org/10.48550/arXiv.2207.04672>. doi:10.48550/ARXIV.2207.04672.
- [24] W. Zhu, Y. Lv, Q. Dong, F. Yuan, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Extrapolating large language models to non-english by aligning languages, CoRR abs/2308.04948 (2023). URL: <https://doi.org/10.48550/arXiv.2308.04948>. doi:10.48550/ARXIV.2308.04948.
- [25] A. Santilli, E. Rodolà, Camoscio: An Italian instruction-tuned LLaMA, in: F. Boschetti, G. E. Leboni, B. Magnini, N. Novielli (Eds.), Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proceedings, Venice, Italy, 2023, pp. 385–395. URL: <https://aclanthology.org/2023.clicit-1.46/>.
- [26] Michael, Stambecco: Italian instruction-following LLaMA model, <https://github.com/mchl-labs/stambecco>, 2023.
- [27] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, A. Mattick, OpenAssistant conversations - democratizing large language model alignment, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023. URL: http://papers.nips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html.
- [28] P. H. Martins, J. Alves, P. Fernandes, N. M. Guerreiro, R. Rei, M. A. Farajian, M. Klimaszewski, D. M. Alves, J. Pombal, N. Boizard, M. Faysse, P. Colombo, F. Yvon, B. Haddow, J. G. C. de Souza, A. Birch, A. F. T. Martins, EuroLLM-9B: Technical report, CoRR abs/2506.04079 (2025). URL: <https://doi.org/10.48550/arXiv.2506.04079>. doi:10.48550/ARXIV.2506.04079.
- [29] H. Li, F. Koto, M. Wu, A. F. Aji, T. Baldwin, Bactrian-X : A multilingual replicable instruction-

- following model with low-rank adaptation, CoRR abs/2305.15011 (2023). URL: <https://doi.org/10.48550/arXiv.2305.15011>. doi:10.48550/ARXIV.2305.15011.
- [30] Z. Xu, F. Jiang, L. Niu, Y. Deng, R. Poovendran, Y. Choi, B. Y. Lin, Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing, in: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025, OpenReview.net, 2025. URL: <https://openreview.net/forum?id=Pnk7vMbznK>.
 - [31] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, B. Li, ALERT: A comprehensive benchmark for assessing large language models’ safety through red teaming, CoRR abs/2404.08676 (2024). URL: <https://doi.org/10.48550/arXiv.2404.08676>. doi:10.48550/ARXIV.2404.08676.
 - [32] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the ITALian language: LLaMAntino-3-ANITA, CoRR abs/2405.07101 (2024). URL: <https://doi.org/10.48550/arXiv.2405.07101>. doi:10.48550/ARXIV.2405.07101.
 - [33] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, R. Xin, Free dolly: Introducing the world’s first truly open instruction-tuned LLM, 2023. URL: <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
 - [34] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.
 - [35] E. Federici, M. Ferraretto, Maestrato chat v0.4 beta, <https://huggingface.co/mii-llm/maestrato-chat-v0.4-beta>, 2024. Hugging Face model card.
 - [36] Almayave, Velvet-14B, <https://huggingface.co/Almayave/Velvet-14B>, 2025. Hugging Face model card.
 - [37] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, X. Shen, Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 5085–5109. URL: <https://aclanthology.org/2022.emnlp-main.340/>. doi:10.18653/v1/2022.emnlp-main.340.
 - [38] OpenAI, gpt-oss-120b & gpt-oss-20b model card, CoRR abs/2508.10925 (2025). URL: <https://doi.org/10.48550/arXiv.2508.10925>. doi:10.48550/ARXIV.2508.10925.
 - [39] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 427–431. URL: <https://aclanthology.org/E17-2068/>.
 - [40] R. Rei, N. M. Guerreiro, J. Pombal, J. Alves, M. A. Farajian, P. Teixeira, A. F. T. Martins, TOWER+: bridging generality and translation specialization in multilingual LLMs, in: M. Liakata, V. P. Moreira, J. Zhang, D. Jurgens (Eds.), Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026, Association for Computational Linguistics, 2026, pp. 29614–29635. URL: <https://doi.org/10.18653/v1/2026.acl-long.1366>. doi:10.18653/V1/2026.ACL-LONG.1366.
 - [41] M. Finkelstein, I. Caswell, T. Domhan, J.-T. Peter, J. Juraska, P. Riley, D. Deutsch, G. Kovacs, C. Dilanni, C. Cherry, E. Briakou, E. Nielsen, J. Luo, K. Black, R. Mullins, S. Agrawal, W. Xu, E. Kats, S. Jaskiewicz, M. Freitag, D. Vilar, TranslateGemma technical report, CoRR abs/2601.09012 (2026). URL: <https://doi.org/10.48550/arXiv.2601.09012>. doi:10.48550/ARXIV.2601.09012.
 - [42] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xCOMET: Transparent machine translation evaluation through fine-grained error detection, Transactions of the Association for Computational Linguistics 12 (2024) 979–995. URL: <https://aclanthology.org/2024.tacl-1.54/>.

doi:10.1162/tac1_a_00683.

- [43] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, LlamaFactory: Unified efficient fine-tuning of 100+ language models, in: Y. Cao, Y. Feng, D. Xiong (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 400–410. URL: <https://doi.org/10.18653/v1/2024.acl-demos.38>. doi:10.18653/v1/2024.ACL-DEMOS.38.
- [44] R. Y. Aminabadi, S. Rajbhandari, A. A. Awan, C. Li, D. Li, E. Zheng, O. Ruwase, S. Smith, M. Zhang, J. Rasley, Y. He, DeepSpeed- inference: Enabling efficient inference of transformer models at unprecedented scale, in: F. Wolf, S. Shende, C. Culhane, S. R. Alam, H. Jagode (Eds.), SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, November 13-18, 2022, IEEE, 2022, pp. 46:1–46:15. URL: <https://doi.org/10.1109/SC41404.2022.00051>. doi:10.1109/SC41404.2022.00051.
- [45] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, HellaSwag: Can a machine really finish your sentence?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4791–4800. URL: <https://aclanthology.org/P19-1472/>. doi:10.18653/v1/P19-1472.
- [46] Y. Bisk, R. Zellers, R. L. Bras, J. Gao, Y. Choi, PIQA: reasoning about physical commonsense in natural language, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 7432–7439. URL: <https://doi.org/10.1609/aaai.v34i05.6239>. doi:10.1609/AAAI.V34I05.6239.
- [47] S. Auer, D. A. C. Barone, C. Bartz, E. G. Cortes, M. Y. Jaradeh, O. Karras, M. Koubarakis, D. Mourmstev, D. Pliukhin, D. Radyush, I. Shilin, M. Stocker, E. Tsalapati, The sciqa scientific question answering benchmark for scholarly knowledge, Scientific Reports 13 (2023) 7240. URL: <https://doi.org/10.1038/s41598-023-33607-z>. doi:10.1038/s41598-023-33607-z.
- [48] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try ARC, the AI2 reasoning challenge, CoRR abs/1803.05457 (2018). URL: <http://arxiv.org/abs/1803.05457>.
- [49] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: <https://openreview.net/forum?id=d7KBjml3GmQ>.
- [50] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, L. Hou, Instruction-following evaluation for large language models, CoRR abs/2311.07911 (2023). URL: <https://doi.org/10.48550/arXiv.2311.07911>. doi:10.48550/ARXIV.2311.07911.
- [51] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the MATH dataset, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021. URL: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>.
- [52] F. M. Molfese, L. Moroni, L. Gioffré, A. Scirè, S. Conia, R. Navigli, Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 18477–18494. URL: <https://aclanthology.org/2025.findings-acl.950/>. doi:10.18653/v1/2025.findings-acl.950.
- [53] L. Gioffré, L. Moroni, A. Fernández-Castro, E. Marafatto, G. Garufi, R. Navigli, INDAQA2 - a large Italian narrative QA benchmark: A CALAMITA 2026 challenge, in: F. Cutugno, A. Miaschi,

- A. Palmero Aprosio, G. Rambelli, L. Siciliani, M. A. Stranisci (Eds.), Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), volume 4195 of *CEUR Workshop Proceedings*, CEUR-WS.org, Bari, Italy, 2026. URL: <https://ceur-ws.org/Vol-4195/69.pdf>.
- [54] E. Kamalloo, N. Dziri, C. Clarke, D. Rafiei, Evaluating open-domain question answering in the era of large language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5591–5606. URL: <https://aclanthology.org/2023.acl-long.307/>. doi:10.18653/v1/2023.acl-long.307.
- [55] T. Bonomo, L. Gioffré, R. Navigli, LiteraryQA: Towards effective evaluation of long-document narrative QA, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 34086–34107. URL: <https://aclanthology.org/2025.emnlp-main.1729/>. doi:10.18653/v1/2025.emnlp-main.1729.
- [56] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, O. Levy, LIMA: Less is more for alignment, in: Thirty-seventh Conference on Neural Information Processing Systems, 2023. URL: <https://openreview.net/forum?id=KBMOKmX2he>.
- [57] Y. Wang, H. Ivison, P. Dasigi, J. Hessel, T. Khot, K. Chandu, D. Wadden, K. MacMillan, N. A. Smith, I. Beltagy, H. Hajishirzi, How far can camels go? exploring the state of instruction tuning on open resources, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: <https://openreview.net/forum?id=w4zZNC4ZaV>.
- [58] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, in: F. M. Nardini, N. Tonellotto, G. Faggioli, A. Ferrara (Eds.), Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023, volume 3448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 9–17. URL: <https://ceur-ws.org/Vol-3448/paper-24.pdf>.

A. Comparison with Existing Italian Instruction Resources

In this Section, we compare DROMEDARIO 3 against the *resources* used to train other Italian-specialized models.

We note that most Italian instruction data was released as a by-product of a model rather than as a documented dataset, and that composition is often reported only in part or not at all. Table 3 summarizes the landscape.

Community-Driven Efforts A first group is published directly as datasets by the Italian open-source community. These include Italian translations of UltraFeedback in both SFT and preference form, Usenet conversation archives, and synthetically generated dialogue such as *capybara-claude-15k-ita*, produced by translating seed prompts and generating turns with proprietary models. The largest is *OpenItalianData*, an Apache-2.0 collection of roughly 1.9M Italian conversations derived from *OpenOrca*. These resources are openly licensed, but each inherits the task profile of the single English corpus it derives from, and the largest of them is undocumented.

Multilingual Corpora with an Italian Slice A second group contains Italian as one language among many. *OpenAssistant Conversations* [27] is among the most reused sources for Italian instruction tuning, across both its OASST-1 and OASST-2 releases. Unlike the other resources in this group, its conversation trees are written by contributors directly in each language rather than machine-translated, so the Italian portion carries no translation artifacts. It is, however, small and restricted to open-ended assistant dialogue, with no coverage of verifiable constraints, mathematics, or code, and its multi-turn trees require flattening into single conversations before use. *Bactrian-X* [29] provides 67K instances per language for 52 languages, with instructions machine-translated from *Alpaca* and *Dolly* and responses generated by GPT-3.5-turbo. The *Aya Dataset* [17] is human-authored but contributes fewer than a thousand Italian instances, while the *Aya Collection* reaches 4.5M, composed mainly of 19 machine-translated English datasets, eleven of which are question answering. Inspecting the Italian split, we observe that NLLB frequently corrupts formatted input such as enumerated answers, a failure mode we do not observe with *TranslateGemma*. *EuroBlocks* [28] is the broadest of these in domain terms, spanning general, coding, mathematical, and STEM instructions across all official EU languages with reward-model and LLM-judge filtering. Its English portion absorbs TüLU 3 wholesale, but Italian content is drawn from unrelated sources and accounts for roughly 8% of the multilingual portion, itself 20% of the 10.6M examples. Domain and language composition are reported separately, so it is not possible to establish from the release which Italian instances, if any, cover code, mathematics, or verifiable constraints.

Resources Released alongside Models A third and largest group exists as the training data of a released model. Some of these datasets are public, others only described, and several not disclosed at all. The earliest are single-corpus translations of the same 52K *Alpaca* instructions: *Camoscio* [25] translates the original *Stanford Alpaca* with GPT-3.5-turbo, *Stambecco* [26] translates the *Alpaca-Cleaned* and GPT-4-LLM variants, and *Fauno* [58] translates the *Baize* sources, using GPT-3.5 for *Alpaca* and *StackOverflow* and Google Translate for the remainder. Two of the three already recognize that some content resists translation: *Stambecco* skips segments detected as code, and *Fauno* selects a stronger engine for the sources known to contain it. Both interventions target a single content type identified in advance, rather than characterizing which instruction categories transfer across languages.

The Italian portion of the *Minerva v1.0* instruction-tuning mixture is the best documented case, assembling roughly 206K instances from ten sources, the largest being *Bactrian-X*, an in-house *Magpie* translation, and an *ALERT*-derived safety set. The subsequent continual-pretraining study [11] takes the opposite balance, shifting its Italian budget from the instruction stage into continual pretraining: the SFT mixture pairs English TüLU 3 with a ~ 30 K Italian supplement of native, translated, and synthetic data, and is applied unchanged to every pretraining variant explored there. *Maestrale* reports scaling from

Resource name	Year	IT size	Chat	Code	Math	IF	Open
<i>Community-driven efforts</i>							
capbara-claude-15k-ita	2024	15K	✓	–	–	–	✓
ultrafeedback-ita	2024	130K	✓	◦	◦	–	✓
OpenItalianData	2026	1.9M [†]	✓	–	◦	–	✓
<i>Multilingual, Italian slice</i>							
Bactrian-X (it)	2023	67K	✓	–	–	–	✓
OASST-2 (it)	2024	<1K	✓	◦	–	–	✓
Aya Dataset (it)	2024	<1K	✓	–	–	–	✓
Aya Collection (it)	2024	4.5M	✓	–	–	–	✓
EuroBlocks (it)	2026	170K [†]	✓	?	?	?	✓
<i>Resources released alongside models</i>							
Alpaca translations (Camoscio/Stambecco)	2023	52K	✓	–	–	–	✓
Fauno	2023	100K [†]	✓	◦	–	–	✓
Minerva v1.0	2024	206K	✓	–	–	–	✓
Maestrale	2024	1.7M [‡]	✓	?	?	?	–
ANITA	2024	100K [✕]	?	?	?	?	✓
Minerva v1.5	2025	30K	✓	–	–	–	✓
Velvet	2025	2M [‡]	?	?	?	?	–
DROMEDARIO 3	2026	630K	✓	✓	✓	✓	✓

Table 3

Italian instruction resources by scale, task coverage, and availability. ✓ present, ◦ partial, – absent, ? undisclosed. Sizes marked [†] are approximate or inferred; those marked [‡] are reported by the authors but with undocumented composition. ANITA[✕] reports no Italian instruction-response pairs, using filtered web text for its adaptation stage instead. For the task coverage, we chose to display four of the most popular broad categories of tasks.

approximately 270K to 1.7M conversations across releases without documenting the composition of the latter. On the industrial side, Velvet reports over 2M SFT examples and over 50K safety instructions, none of which are released.

Positioning DROMEDARIO 3 differs from all of the above along three axes. First, *provenance*: it is derived from a single documented English mixture with per-instance source attribution, which makes it the only Italian instruction resource with a matched English counterpart of this kind. This property is what makes the mixing study of Section 5 possible, since it allows English and Italian conditions to differ in language while holding composition fixed. Second, *composition*: precise instruction following, mathematics, code, and safety are absent from every prior Italian resource or present only as an undocumented fraction of a closed mixture, whereas DROMEDARIO 3 inherits the full domain coverage of its source and characterizes each domain by transferability. Third, *intended use*: prior resources are released as adjuncts to a single model, without ablations over how much target-language data is needed. DROMEDARIO 3 is released as a standalone resource together with a controlled study of English-Italian mixing ratios, and is intended to be a practical guide for researchers interested in Italian.

B. Translatability of Source Datasets

This Appendix provides supplementary details on the source-dataset filtering step described in Section 3.1. Table 4 lists the five datasets that were either excluded from the translation pipeline or retained in their original form, together with the rationale for each decision and a representative prompt example.

C. Taxonomy-Based Classification Prompts

As described in Section 3.1, we use GPT-oss-120b to assign each instance to a task type and a domain from the SUPER-NATURALINSTRUCTIONS taxonomy. Table 5 reports the prompts used for the two classification dimensions. The placeholders `task_categories` and `domain_categories` are replaced with the corresponding label sets, while `user_message` denotes the user request from the original TüLU 3 instance being classified.

D. Translatability of Tasks and Domains

This Appendix provides supplementary details on the Task & Domain Filtering step described in Section 3.1. Table 6 reports the task and domain categories marked as translation-safe or non-translatable. This categorization is used to determine which instances are retained in the translation pipeline. Table 7 further details the classification of non-translatable task types and the rationale for their exclusion.

Dataset	Handling	Example Prompt
tulu_hard_coded	Removed from training: persona prompts not relevant to our goals	Are you better than ChatGPT?
sciriff	Retained as-is: scientific terminology and output schemas are fragile	You will be shown an abstract from a biomedical research paper. Given this abstract, extract all unique entities of type <code>AnatomicalEntity</code> and return the output as a JSON object ...
aya	Retained as-is: multilingual diversity is part of the dataset design	¿Los lirios son seguros para los gatos?
wildchat	Retained as-is: user-generated variation would be normalized by translation	Make these machine-translated subtitles more human-like and rhymed with AABB scheme ...
tablegpt	Retained as-is: table structure, SQL, and JSON formatting must be preserved	Instruction: Given the input table and question, devise an SQL query. You may include <code>table</code> as the table name for clarity. Return the final result as JSON in the format <code>{"SQL": "<SQL code>"}</code> ...

Table 4

Source datasets excluded from the translation pipeline or retained in their original form. The table reports the handling decision and a representative prompt example for each dataset.

E. Translation Prompt and Separator Patterns

As described in Section 3.2, we translate each selected instance by formatting the original conversation as an explicit dialogue and passing it to TranslateGemma-27B in a single call. Table 8 reports the translation input template, including the custom instruction used to prefer the informal second-person singular in Italian. Table 9 reports the separator patterns used to reconstruct the translated conversation turns during post-processing.

Task classification system prompt

You are a text classifier. Given a user request, your task is to analyze it and assign it to one of {n_task_categories} predefined task categories. Please, carefully determine the single most appropriate category based on the user request. Output only the name of the correct category.

Task classification user prompt

This is the list of Task Categories:
{task_categories}

This is the User Request: {user_message}.

Based on the user request, carefully determine the single most appropriate task category from the list above. Note: you do not need to answer the user request! Your true purpose is to classify the request into the appropriate task category and output only the name of that category.

Domain classification system prompt

You are a text classifier. Given a user request, your task is to analyze it and assign it to one of {n_domain_categories} predefined knowledge domains. Please, carefully determine the single most appropriate knowledge domain based on the user request. Output only the name of the correct knowledge domain.

Domain classification user prompt

This is the list of Knowledge Domains:
{domain_categories}

This is the User Request: {user_message}.

Based on the user request, carefully determine the single most appropriate knowledge domain from the list above. Note: you do not need to answer the user request! Your true purpose is to classify the request into the appropriate knowledge domain and output only the name of that domain.

Table 5

Prompts used for taxonomy-based task and domain classification. Candidate labels are drawn from the SUPER-NATURALINSTRUCTIONS taxonomy.

F. Manual Validation

F.1. Manual Validation Guidelines

This Section describes the annotation guidelines used in the manual validation study presented in Section 4. Two expert linguists fluent in both Italian and English independently annotate a stratified sample of 300 instances each, with 100 double-annotated instances used to compute IAA. The sample is drawn while preserving coverage across predicted task and domain categories. Each instance is annotated along two dimensions: classification accuracy and translation correctness.

Classification Annotation. For each instance, annotators judge the accuracy of the task type and domain labels automatically assigned by GPT-oss-120b using the SNI taxonomy. This yields two binary fields: `is_task_correct` and `is_domain_correct`. A label is marked as correct if it represents the most appropriate category for the given user prompt; annotators are provided with the full SNI label set as reference.

Type	Translation-safe	Non-Translatable
Tasks	Answer Verification, Answerability Classification, Cause Effect Classification, Code to Text, Coherence Classification, Commonsense Classification, Data to Text, Dialogue Act Recognition, Dialogue Generation, Ethics Classification, Explanation, Fact Verification, Information Extraction, Keyword Tagging, Mathematics, Misc., Negotiation Strategy Detection, Program Execution, Question Answering, Question Generation, Question Understanding, Section Classification, Sentence Composition, Sentence Compression, Sentence Ordering, Sentiment Analysis, Spam Classification, Stereotype Detection, Story Composition, Summarization, Text Categorization, Text Completion, Text to Code, Textual Entailment, Title Generation, Wrong Candidate Generation	Coreference Resolution, Dialogue State Tracking, Discourse Connective Identification, Discourse Relation Classification, Entity Generation, Entity Relation Classification, Fill in The Blank, Gender Classification, Grammar Error Correction, Grammar Error Detection, Intent Identification, Irony Detection, Language Identification, Linguistic Probing, Named Entity Recognition, Number Conversion, Overlap Extraction, POS Tagging, Paper Review, Paraphrasing, Poem Generation, Preposition Prediction, Punctuation Error Detection, Question Decomposition, Question Rewriting, Sentence Expansion, Sentence Perturbation, Speaker Identification, Speaker Relation Classification, Spelling Error Detection, Stance Detection, Style Transfer, Text Matching, Text Quality Evaluation, Text Simplification, Toxic Language Detection, Translation, Word Analogy, Word Relation Classification, Word Semantics
Domains	Accounting, Animals, Anthropology, Architecture, Art, Astronomy, Biology, Books, Business Ethics, Captions, Chemistry, Code, Commonsense, Computer Science, Conference, Countries, Debatepedia, Dialogue, Econometrics, Economics, Electrical Engineering, Fiction, Food, Formal Fallacy, Game, Geography, Global Facts, Government and Politics, Healthcare, History, Human Race, Human Sexuality, International Law, Jurisprudence, Justice, Knowledge Base, Law, Literature, Logic, Management, Marketing, Mathematics, Medical Genetics, Medicine, Miscellaneous, Moral Scenarios, Movies, Music, Narrative, Natural Science, News, Nutrition, Personal Narratives, Philosophy, Physics, Pop Culture, Prehistory, Professions, Psychology, Public Places, Public Relations, Reviews, Scientific Research Papers, Security: Environmental Security, Security: National Security, Social Media, Sociology, Sports, Statistics, Stereotypes, Story, TED Talks, US Foreign Policy, Web, Wikipedia, World Religions	English Exams, Linguistics

Table 6

Breakdown of translation-safe and non-translatable task types and domains.

Translation Annotation. For each instance, annotators assess the quality of the Italian translation produced by TranslateGemma-27B. The binary field `is_translation_correct` records the overall verdict. Importantly, annotators are instructed to apply a non-perfectionistic standard: a translation is considered correct if it reads naturally in Italian and preserves the intent and task structure of the original English instance, even if it is not a verbatim or maximally faithful rendering. Minor stylistic variation, paraphrasing, or slight rephrasing is therefore not penalized, provided the translated instance

Category	Excluded task types	Rationale
Language-form dependent tasks	Discourse Connective Identification, Grammar Error Correction, Grammar Error Detection, Linguistic Probing, Overlap Extraction, POS Tagging, Poem Generation, Preposition Prediction, Punctuation Error Detection, Spelling Error Detection, Word Analogy	These tasks depend on surface-level properties of the source language, such as morphology, spelling, punctuation, word order, lexical form, or language-specific discourse markers. Translating them would alter or remove the phenomena the task is intended to test.
Label-unstable or meaning-sensitive tasks	Coreference Resolution, Gender Classification, Irony Detection, Number Conversion, Paraphrasing, Question Rewriting, Stance Detection, Text Quality Evaluation, Word Relation Classification, Word Semantics	These tasks rely on fine-grained semantic, pragmatic, cultural, or annotation-specific distinctions that may not be preserved under translation. As a result, labels or expected outputs may become ambiguous, invalid, or no longer aligned with the translated input.
Structure- or format-sensitive tasks	Dialogue State Tracking, Discourse Relation Classification, Entity Relation Classification, Fill in The Blank, Intent Identification, Question Decomposition, Sentence Perturbation, Style Transfer, Text Matching, Text Simplification	These tasks involve placeholders, special tokens, span alignments, structured formats, or controlled transformations. Machine translation may alter, drop, or reorder these elements, making reconstruction and evaluation unreliable.
Methodologically invalid after translation	Language Identification, Toxic Language Detection, Translation	For these tasks, translation either makes the task circular, trivial, or unreliable.
Low-value or high-risk translation cases	Entity Generation, Named Entity Recognition, Paper Review, Sentence Expansion, Speaker Identification, Speaker Relation Classification	For these tasks, translation is either unnecessary, weakly informative, or likely to compromise annotation quality. Speaker cues, entity spans, long contexts, and review-style judgments can be distorted by translation or are more meaningful in the original language.

Table 7

Classification of non-translatable task types, grouped by the main rationale for their exclusion from the machine translation pipeline.

remains semantically equivalent and well-formed as an instruction.

When a translation is judged incorrect, annotators additionally assign an error type using the `translation_error_type` field. Four categories are defined: (1) *incorrect translation*, covering cases where the output is substantially wrong in meaning or content; (2) *incomplete translation*, covering cases where the output is truncated or where part of the instance is left in the source language; (3) *task execution*, covering cases where the model produces a response to the prompt instead of a translation of it; and (4) *other*, used for recurring or novel failure patterns not captured by the preceding categories.

F.2. Translation Error Examples

Table 10 reports representative translation errors identified during the manual validation described in Section 4, grouped by the error type assigned by the annotators.

TranslateGemma input template

«<custom>» Use the informal second person singular instead of the formal or plural form.

«<source>>en«<target>>it-IT«<text>>

User1 asks:

{user_message}

User2 replies:

{assistant_message}

Generation prefill

L'utente1 chiede:

Table 8

Translation input template used for TranslateGemma-27B. `user_message` and `assistant_message` are replaced with the user and assistant turns from the original TULU 3 instance. The generation prefill is used to steer the model toward translation rather than task execution.

Turn type	Accepted separator patterns
User turn	L'utente1 chiede:
	L'utente1 domanda:
	L'utente1 dice:
	L'utente1 scrive:
Assistant turn	L'utente2 risponde:
	L'utente2 replica:
	L'utente2 dice:
	L'utente2 scrive:
	User2 replies:

Table 9

Separator patterns used to split translated outputs and reconstruct the original conversation structure. Patterns are matched case-insensitively and allow minor spacing and apostrophe variants.

Error Type	Source (English)	Generated Translation (Italian)
<i>Incorrect Translation</i>	...the most frequently <i>recommended</i> literary work ...	...l'opera letteraria più frequentemente <i>raccomandata</i> ...
<i>Incorrect Translation</i>	...evaluate the sentiment and <i>emotional tone</i> ...	...valutare il sentimento <i>and emotional tone</i> ...
<i>Incomplete Translation</i>	...Step 5: Determine the coordinates for each text block. ...Step 6: Summarize the solution. ...	...Passaggio 5: determina le coordinate per ciascun blocco di testo. ...[TRUNCATION]

Table 10

Representative translation errors identified during manual validation, grouped by the error type assigned by the annotators. The examples show a semantic calque (*recommended* → *raccomandata* instead of *consigliata*), an instance only partially translated into Italian, and a truncated output.

G. Training Details

We report the configuration for all training runs described in Section 5 in Table 11. We adopt the same hyperparameters as Lambert et al. [4], except for effective batch size, which we set to 192 instead of 128 to better optimize computation and communication on our machines.

Hyperparameter	Value
Epochs	2
Effective batch size	192
Mini batch size	6
Gradient accumulation steps	2
Number of nodes	4
Number of GPUs per node	4
Sequence length	4096
Learning rate	5.0e-6
Learning rate scheduler	linear
Warmup ratio	0.03

Table 11

Hyperparameters used in all our training runs. We follow those of Lambert et al. [4] except for the mini batch size, which we raise from 4 to 6 to optimize computation on our infrastructure.