Reasoning-Centric Cross-Lingual Alignment for Low-Resource Dialect Mastery with Minimal Data: A Darija Case Study

Anonymous ACL submission

Abstract

The dominance of high-resource languages in open-source large language models (LLMs) threatens digital sovereignty and linguistic inclusion. Bridging this gap typically requires costly fine-tuning, often leading to catastrophic forgetting of multilingual abilities.

We propose a data-efficient, reasoning-centric alignment approach for adapting open LLMs to under-resourced dialects. Applied to Moroccan Arabic (Darija), our method combines the translation of compact, high-quality instruction datasets with selective English retention and parameter-efficient fine-tuning using *LoRA*. The resulting GEMMAROC-27B model achieves parity with ATLAS-CHAT-27B on *DarijaMMLU* (61.6%) and notable gains on Darija *HellaSwag* (62.3% vs. 48.4%), while showing only moderate trade-offs on English reasoning benchmarks such as *GSM8K* and *MMLU*. Gains are consistent across reasoning tasks, though performance on other Darija applications (e.g., sentiment analysis) remains more variable due to limited task-specific data.

Replication on QWEN-2.5 confirms the model-agnostic nature of the recipe. Fine-tuning GEMMAROC-27B required only **48 GPU·hours**, illustrating that strong dialectal alignment can be achieved sustainably and at low computational cost.

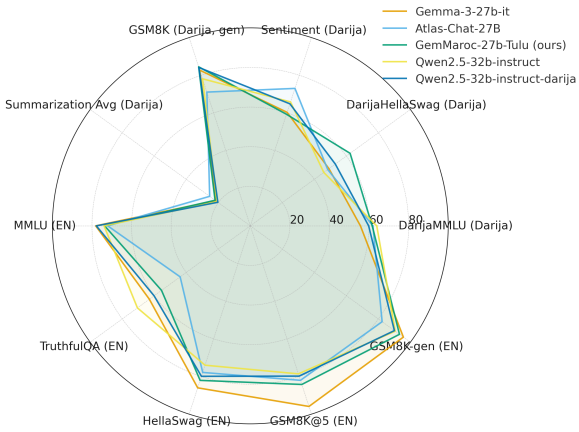

Figure 1: Model performance across Darija and English benchmarks. Curves show GEMMAROC-27B-TULU alongside GEMMA-3-27B-IT, ATLAS-CHAT-27B, QWEN2.5-32B-INSTRUCT, and QWEN2.5-32B-INSTRUCT-DARIJA. Higher is better.

1 Introduction

Large Language Models (LLMs) are increasingly central to AI development, but their progress remains concentrated in a handful of high-resource languages (Qin et al., 2025) (English, Mandarin, etc.), creating strategic and linguistic inequities. This imbalance particularly affects languages such as Moroccan Arabic (Darija), spoken by over *40 million* people, yet largely absent from open-source AI ecosystems. This exclusion undermines digital inclusion and poses a major challenge to Morocco's national digital transformation goals outlined in *Morocco Digital 2030* (MDTAR, 2024).

Two common routes to bridge this gap, training new large models from scratch or fine-tuning existing LLMs, carry limitations. Training from scratch is prohibitively expensive for most institutions. Naive fine-tuning requires large, high-quality instruction sets and risks catastrophic forgetting of previously acquired multilingual and reasoning competencies. Prior Darija efforts, such as ATLAS-CHAT (Shang et al., 2025), demonstrate feasibility but rely on hundreds of thousands of examples, raising questions of scalability and environmental cost.

We present a *data-efficient, reasoning-centric alignment* method to adapt open LLMs to under-resourced dialects while preserving general abilities. Compact, high-quality instruction data are translated into the target dialect, with a controlled share of English examples to prevent knowledge drift. Parameter-efficient fine-tuning with LoRA minimizes compute and energy use, consistent with

Green AI principles. Although these components have been studied separately, their joint use and evaluation in low-resource dialect adaptation remain unexplored.

We also instantiate this methodology in GEMMAROC, a family of Darija models based on GEMMA-3. The resulting GEMMAROC-27B achieves parity with ATLAS-CHAT-27B on *DarijaMMLU* (61.6 %) and notable gains on *Darija HellaSwag* (62.3 % vs. 48.4 %), with only moderate trade-offs on English reasoning benchmarks such as *GSM8K* and *MMLU*. Replication on QWEN-2.5 confirms the model-agnostic nature of the recipe, and the entire fine-tuning of GEMMAROC-27B required only **48 GPU·hours**, highlighting the potential of reasoning-centric alignment for sustainable and inclusive multilingual LLM development. Results are summarized in a unified leaderboard: Figure 1 for model performance and Table 1 for Darija and English tracks.

Thus, the main contributions of the paper are:

- **Minimal-data alignment methodology.** We propose a data-efficient recipe for extending open-source LLMs to low-resource dialects, demonstrated on Darija, with concrete guidelines for data selection, language balance, and reasoning integration.
- **Darija resources and GEMMAROC models.** We release a lean instruction suite (translated *LIMA*, *DEITA*, and reasoning-rich *TULU*) along with our GEMMAROC-27 model that achieve state-of-the-art Darija performance in reasoning tasks at minimal cost.
- **Evaluation and openness.** We introduce new benchmarks, including *Darija GSM8K*, and conduct comprehensive human evaluations confirming GEMMAROC's fluency and reasoning ability. All models, code, datasets, and benchmarks are openly released.[1]

2 Related Work

Developing effective LLMs for low-resource dialects like Darija intersects three key research areas: (1) diagnosing limitations of existing Arabic LLMs, (2) advancing data-efficient and cross-lingual alignment, and (3) integrating reasoning into multilingual models.

[1]Full code is available at https://github.com/abdelnlp2/eaclanonyme, with data and models at https://huggingface.co/MathematicianNLPer.

2.1 Darija and Arabic LLMs

Most existing Arabic LLMs focus on Modern Standard Arabic (MSA), demanding vast computational resources while neglecting dialectal diversity. Models like ALLAM (Bari et al., 2025) and FANAR (Team et al., 2025) perform well thanks to large-scale training on hundreds of billions to trillions of tokens, but their reliance on heavy compute and massive datasets limits reproducibility and accessibility, making them impractical for low-resource or dialect-focused adaptation.

Conversely, ATLAS-CHAT (Shang et al., 2025) pioneered Darija-specific fine-tuning with 458K native and translated instructions, achieving strong *DarijaMMLU* results but limited generalization due to its monolingual setup. Building on this, AL-ATLAS (Bounhar and Majjodi, 2025) demonstrated that small, dialect-pretrained models (0.5B) can outperform larger systems in human evaluation. Meanwhile, Cohere's COMMAND R+ (Cohere et al., 2025) achieved efficient multilingual adaptation through SRPO and hybrid experts, attaining 68.8% on *NTREX* without full Arabic retraining, offering key insights for resource-efficient dialectal fine-tuning.

2.2 Data-Efficient and Cross-Lingual Alignment

Research increasingly demonstrates that building effective dialectal models requires a paradigm shift from emphasizing data volume to prioritizing data quality, curation, and cross-lingual transfer. The *Superficial Alignment Hypothesis* (Zhou et al., 2023) suggests that pretrained LLMs primarily benefit from supervised fine-tuning (SFT) for stylistic and alignment purposes. Consequently, small but high-quality datasets, sometimes as few as 1k examples, can rival models trained on much larger yet less diverse corpora.

Methods like *DEITA* (Liu et al., 2024) and *MIG* (Chen et al., 2025) show that selecting only the most informative 5–20% of data—based on linguistic complexity or semantic diversity—can match or surpass full-data performance, streamlining SFT. Likewise, *DEFT-UCS* (Das and Khetan, 2024) and *UniMax* (Han et al., 2025) demonstrate that advanced selection can achieve full-data parity with just 10–32% of the dataset.

For low-resource languages, cross-lingual strategies are essential. *Parallel* instruction tuning with English and the target language (Weber et al., 2024)

improves accuracy by up to +9.9 over monolingual tuning. Pivot-based methods like *PLUG* (Zhang et al., 2024) achieve strong results with only 2k samples. Research on cross-lingual alignment (Li et al., 2024), mixed English–native prompts, and scalable synthetic frameworks such as *Bactrian-X* (Li et al., 2023) and *MURI* (Köksal et al., 2024) highlights the power of automated translation pipelines (Holmström and Doostmohammadi, 2023) for building instruction datasets efficiently.

2.3 The Role of Reasoning in Alignment

Recent work on multilingual reasoning (Chain-of-Thought, CoT) shows that translating high-quality reasoning datasets significantly boosts performance. *MathOctopus* (Chen et al., 2024) and *XCOT* (Chai et al., 2024) report gains of up to 25 points on MGSM using translated and code-switched examples. Similarly, *mCoT* (Lai and Nissim, 2024) built a 6.3M-example corpus from MetaMathQA translations, where fine-tuning on translated CoT data enabled 7B models to outperform larger proprietary systems (e.g., GPT-3). These results validate translation-based CoT as an efficient path to improved dialectal reasoning, motivating our use of the TULU dataset for Darija fine-tuning.

3 Reasoning-Centric Cross-Lingual Alignment

Building upon recent advances in low-resource language alignment, we propose a highly efficient and ecologically sustainable fine-tuning methodology to unlock latent Moroccan Darija capabilities in LLMs (Figure 2).

3.1 Frontier Open-Source Backbone Models

The methodology starts from a frontier open-source model such as GEMMA-3 or QWEN-2.5, which already shows strong multilingual and reasoning abilities. Though not trained on Moroccan Darija, these models exhibit latent competence from related Arabic dialects. Leveraging this foundation, rather than training from scratch, ensures both efficiency and sustainability.

3.2 Efficient Data Selection

At the core of our approach lies the selection of SFT data, informed by insights from our Related Work section (see Section 2), particularly studies on data-efficient fine-tuning (Das and Khetan, 2024; Zhou et al., 2023; Liu et al., 2024; Chen et al., 2025). Only a compact but diverse and high-quality subset of data is selected, focusing on:

- **CoT samples:** Following multilingual reasoning works such as *XCOT* and *mCoT*, we prioritize data with explicit reasoning traces and translation-based CoT to enhance both reasoning and cross-lingual generalization.
- **Consistent and high-quality data:** In line with *LIMA* and *DEITA*, we select a small yet carefully curated set of instruction-response pairs emphasizing coherence, informativeness, and stylistic alignment.
- **Diverse domains:** Inspired by *DEFT* and *MIG*, we ensure diverse coverage across reasoning, dialogue, and cultural domains to foster generalization and avoid overfitting.

3.3 Data Preparation and Translation

Selected data are then processed to align with Darija while retaining cross-lingual grounding. The process involves:

- Filtering out all multilingual data, retaining only English instructions: This prevents errors from prompts mentioning specific languages (e.g., "Translate from language A to B" or "Answer in language X"), which, if translated, would compromise data quality. For accurate filtering, an XLM-Roberta-based detector can be used (ProtectAI, 2025). To preserve technical content, code, LaTeX formulas, and specialized terms need to be kept in their original form, translating only when clear Arabic equivalents exist (e.g., معادلة for equation, دالة for function). This ensures consistency with standard scientific usage.
- Translating high-quality instruction and reasoning samples into Moroccan Darija: Translations can be generated with any online translation API using a tailored prompt for Darija in Arabic script, defaulting to Modern Standard Arabic when uncertain. An example of a full prompt is provided in the Appendix A.1.1.
- Retaining a small fraction of English data: To mitigate catastrophic forgetting (Mao et al., 2022; Wu et al., 2025), we retain about 30% of the data in English across all datasets, preserving cross-lingual transfer and general reasoning ability (Wu et al., 2024; Li et al., 2024; Zhang et al., 2024; Li et al., 2023).

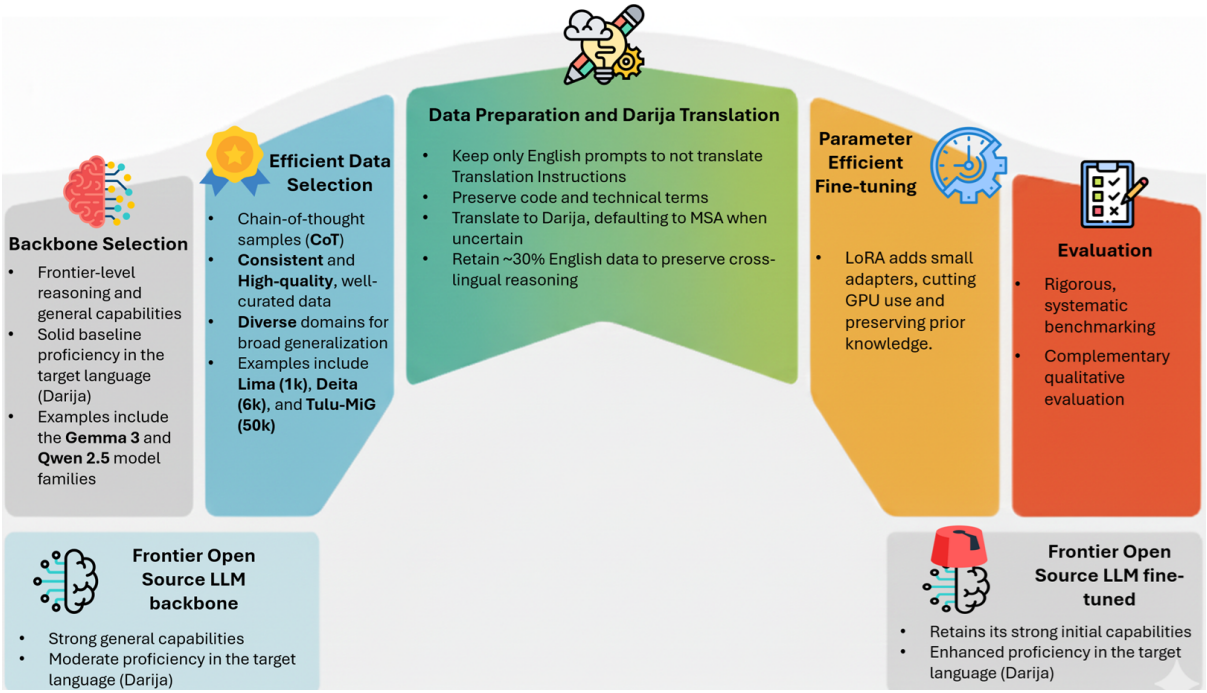

Figure 2: Reasoning-Centric Cross-Lingual Alignment to improve Darija model performance efficiently.

Manual Verification. Translation outputs were periodically sampled and reviewed for quality. Instead of manual corrections, translation prompts were iteratively refined based on native Darija fluency when recurring issues arose. Prompts requiring English responses were left untranslated to preserve meaning. While not flawless, the dataset's quality was deemed sufficient for fine-tuning and consistent with the study's exploratory, low-resource goals.

3.4 Parameter-Efficient Fine-Tuning (LoRA)

Fine-tuning is performed using *LoRA* (Hu et al., 2022), which introduces small trainable adapters while freezing most model parameters. This method drastically reduces GPU and energy consumption. It also mitigates catastrophic forgetting, allowing the model to learn new dialectical knowledge while retaining prior multilingual and reasoning abilities. Multi-turn dialogue data can be formatted following the Vicuna conversational template (Chiang et al., 2023; LMSYS Org, 2023).

3.5 Systematic & Qualitative Evaluation

Before deployment, fine-tuned models undergo a rigorous passgate evaluation combining quantitative benchmarks and qualitative human assessment. The goal is to ensure that the adapted model preserves the original strengths of the frontier open-source backbone, avoiding catastrophic forgetting, while achieving tangible gains in Darija competence. Evaluation spans four quantitative axes: language competence, general reasoning, mathematical reasoning, and truthfulness & safety. In addition, a focused human evaluation assesses Helpfulness & Alignment, Completeness, and Darija Fluency, providing a holistic view of both technical robustness and linguistic naturalness.

4 Experimental Setup

To validate our approach, we run controlled fine-tuning on frontier open-source LLMs, detailing datasets, preprocessing, backbone setup, fine-tuning parameters, and evaluation benchmarks for language, reasoning, and robustness.

4.1 Backbone configurations

We validate our approach by fine-tuning two backbone families: GEMMA-3 (GEMMAROC series) and QWEN-2.5 (used later for ablation). Specifically, we fine-tuned GEMMA-3-4B and GEMMA-3-27B (GEMMAROC 3-27B), chosen for their strong reasoning and multilingual capabilities.

We emphasize that the newer GEMMA-3 series offers a clear advantage over the GEMMA-2 backbone used in ATLAS-CHAT. Our comparison is therefore not meant to assert methodological superiority but to benchmark against a strong SOTA Darija model. The goal is to show that, regardless

of architecture, a minimal-data, reasoning-centric approach can fully leverage modern backbones to achieve high-efficiency, near–SOTA Darija performance. Comparisons with ATLAS-CHAT are thus indicative rather than competitive. The backbone gap (GEMMA-3 vs. GEMMA-2) is acknowledged as a confounding factor, and our focus remains on assessing the effectiveness of the alignment methodology rather than architectural gains.

4.2 Efficient Data Selection

Good candidates for the selection criteria described in 3.2 are the following three compact to mid-scale instruction corpora, each translated into Moroccan Darija, to surface the latent capabilities of multilingual backbones under limited computational budgets (Zhou et al., 2023; Liu et al., 2024; Chen et al., 2025):

- **LIMA 1K** (Zhou et al., 2023): 1,000 single-turn tasks that emphasize stylistic quality and input diversity. Despite its size, LIMA often rivals much larger instruction sets, showing that small, well-curated data can unlock pre-trained knowledge.
- **DEITA 6K** (Liu et al., 2024): 6,000 multi-turn prompts selected with EVOL-COMPLEXITY and EVOL-QUALITY scores. It bolsters dialogue coherence and contextual nuance in Darija.
- **TULU 50K** (Chen et al., 2025): 50,000 information-gain–maximised instructions rich in mathematics, coding, and science domains underrepresented in earlier Darija work, such as Shang et al. (2025).

Comprehensive dataset descriptions appear in Appendix A.1.2, with final sizes after filtering and translation shown in Table 6 (Appendix A.2). To meet model context limits (2,048 tokens), over-length samples are iteratively trimmed—dialogues shortened by removing the last user–assistant turns—discarding only those still exceeding the limit.

4.3 Training Details

Hyperparameters: Training used `bf16` precision. LoRA ranks and scales were r=32, α=64 for the 4B model and r=16, α=32 for the 27B model. Learning rates were 1×10^{-4} for *TULU* and 4×10^{-4} for *DEITA* and *LIMA*. Models trained for 3, 6, and 15 epochs with a 2,048-token limit, following Zhou et al. (2023); Liu et al. (2024); Chen et al. (2025).

Hardware/Cost: All runs used *2×A100-80GB* GPUs, except GEMMAROC-27B on *8×H100-80GB*. The largest run lasted about *6 hours*.

4.4 Evaluation

4.4.1 Baselines

We compare against (i) the *untuned* GEMMA checkpoints (4 B & 27 B), (ii) ATLAS-CHAT (9 B & 27 B) (Shang et al., 2025), the previous state of the art for Darija, and (iii) ALLAM (7 B & 13 B) (Bari et al., 2025), which remains the state of the art for large-scale Arabic instruction tuning.

4.4.2 Systematic Evaluation

We design a comprehensive benchmark suite spanning four dimensions:

Darija competence (Shang et al., 2025)

- *DarijaMMLU*: 22,027 multiple-choice questions across 57 subjects, filtered for Moroccan relevance.
- *DarijaHellaSwag*: 10k adversarial commonsense continuations in Darija.
- *DarijaBench*: Covering Sentiment Analysis and Summarization (Translation and Transliteration dropped for low quality).

Reasoning, World Knowledge, Truthfulness

- *MMLU*: 15,908 multiple-choice questions over 57 subjects (math, law, medicine, history), averaged across subjects (Hendrycks et al., 2021).
- *HellaSwag (EN)*: 10k adversarial common sense continuations (Zellers et al., 2019).
- *TruthfulQA*: 817 short-answer questions across 38 domains targeting common misconceptions (Lin et al., 2022).

Mathematical reasoning

- *GSM8K*: 1,000 grade-school problems (2–8 steps), evaluated in nonchat mode with 5-shot prompting (Cobbe and et al., 2021).
- *GSM8K-gen*: zero-shot chat adaptation prompting for step-by-step reasoning and boxed answers; parsed for numeric result.

- *GSM8K-darija-gen*: Darija version of *GSM8K-gen* introduced in this paper, with problem statements automatically translated from English and manually spot-checked for semantic fidelity. Despite minor translation noise, it constitutes the first publicly available Darija reasoning benchmark and remains suitable for comparative evaluation.

Metrics and post-processing. For *GSM8K@5*, five few-shot examples include CoT reasoning with a final "####" answer requiring an exact numeric match. *GSM8K-gen* is evaluated zero-shot, parsing boxed outputs for the final numeric answer. Task-specific metrics (classification, summarization, etc.) are reported in the results tables.

4.4.3 Qualitative Evaluation

Moroccan law presents a compelling test case: while statutes are written in MSA or French, citizens typically require explanations in Darija. We design a small-scale evaluation to assess whether GEMMAROC-27B can accurately transform legal texts into accessible guidance.

Participants and Profiles. The authors (native speakers) acted as evaluators under three profiles: a law student (precision, statutory grounding), a civil servant/NGO worker (clarity, protections), and an ordinary citizen (understandability, usefulness). Each profile handled 10 prompts (5 Family Law, 5 Labour Law).

Task Design and Scenarios. Both ATLAS-CHAT-27B and GEMMAROC-27B answered 30 prompts, each specifying *persona*, *question*, and *legal excerpt*: 1) **Scenario A: Family Law** (Topics: marriage formalities, judicial age exceptions), and 2) **Scenario B: Labour Law** (Topics: working time, rest, severance, protections).

Weights and Scoring. Outputs were rated on four 1–5 scales for helpfulness and alignment, accuracy to the legal text (favoring cautious answers), completeness (covering all relevant points), and Darija fluency. A weighted composite score was computed per item using $w = \langle 0.12, 0.50, 0.13, 0.25 \rangle$. Accuracy received the highest weight (0.50) as faithfulness to legal content is paramount. Fluency (0.25) was next, since open-source LLM backbones often struggle with natural Darija. The remaining criteria, helpfulness/alignment and completeness, were given smaller, nearly equal weights to reflect their secondary yet supportive roles.

Blind Evaluation Protocol. Evaluation was double-blind: anonymized outputs were independently rated to ensure fairness. Although the sample size is modest, this setup serves as an initial validation. Future iterations will expand the pool of raters and include external Darija speakers to ensure statistical robustness and inter-rater reliability.

5 Results and Discussion

This section presents the empirical results and analysis from Sections 3 and 4.

5.1 Systematic Evaluation Results

Table 1 shows the unified leaderboard combining the Darija- and English-centric benchmarks, comparing our models (marked "ours") against strong baselines across both tracks. To ensure robustness, all reported scores are averaged over two random seeds. Random-seed values, prompt templates, and evaluation scripts are documented to ensure full reproducibility. As most metrics are accuracies expressed in percentages, we report their variability as ± percentage points (pp). One can notice that the variance remains relatively low across the board: *DarijaMMLU* (±1.2 pp), *DarijaHellaSwag* (±2.5 pp), *Sentiment* (±0.9 pp), *GSM8K-darija-gen* (±1.3 pp); and for English benchmarks: *MMLU* (±1.1 pp), *HellaSwag* (±2.3 pp), *GSM8K* (±0.8 pp), *GSM8K-gen* (±1.0 pp), and *TruthfulQA* (±0.5 pp). Figure 1 provides a complementary visualization, highlighting relative strengths and weaknesses across benchmarks in both Darija and English. To keep the figure clear and interpretable, only the top-performing larger models are shown.

Overview. We analyze how fine-tuning datasets and model scales shape the emergence and consolidation of Darija competence. Comparisons are made between instruction-tuned GEMMA3 baselines and Darija-adapted variants fine-tuned on one of three datasets: *Lima* (1K), *Deita* (6K), or *Tulu* (50K). We also examine cross-model trends at larger scales to assess whether the gains generalize across backbones. The goal is to understand how instruction tuning and reasoning-centric alignment foster dialectal specialization while preserving cross-lingual reasoning.

Effect of Data Scale (4B models). All three 4B variants exhibit measurable adaptation effects, but

Language	Darija									English				
Model/Benchmark	Size (B)	DarijaMMLU	DarijaHellaSwag	Sentiment Analysis	GSM8K-darija-gen	Summarization				MMLU	TruthfulQA	HellaSwag	GSM8K @5	GSM8K-gen
	Metrics	Acc	Acc	Acc	Acc	chrF	ROUGE-1	ROUGE-L	BERTScore	Acc	bleu_acc	Acc	Acc	Acc
GEMMA3-4B-IT	4	32.8	36.3	58.94	52.77	27.22	8.38	8.19	37.23	51.1	40.88	47.65	74.75	89.08
GEMMAROC-4B-LIMA (OURS)	4	34.95	39.27	42.1	30.05	26.14	6.95	7.04	34.32	29.28	40.15	44.21	51.24	65.42
GEMMAROC-4B-DEITA (OURS)	4	42.67	44.26	60.8	34.12	27.16	7.4	7.34	38.48	51.35	44.55	68.97	53.15	61.64
GEMMAROC-4B-TULU (OURS)	4	47.53	47.13	53.29	37.91	28.46	8.89	8.76	39.27	54.14	43.33	73.95	55.95	71.57
ALLAM-INSTRUCT-7B	7	59.49	50.09	47.33	40.33	10.27	1.68	1.68	12.28	58.31	42.11	75.2	49.28	68.61
ATLAS-CHAT-9B	9	58.32	43.65	81.85	66.69	32.07	9.5	9.45	47	69.09	**67.56**	73.35	73.01	77.03
ATLAS-CHAT-27B	27	**61.95**	48.37	**73**	71.04	**32.75**	**10.53**	10.42	**47.82**	72.06	43.82	77.84	82.03	82.34
GEMMA-3-27B-IT	27	55.65	49.13	60.27	82.71	28.33	10.28	9.95	38.17	**78.12**	**63.05**	**86.02**	**95.9**	**95.60**
GEMMAROC-27B-TULU (OURS)	27	**61.61**	**62.33**	59.25	**84.46**	28.34	9	**11.2**	**39.5**	**73.6**	55.45	**82.06**	**84.23**	**93.18**

Table 1: Unified leaderboard comprising: (1) Darija-centric benchmarks, and (2) English-centric benchmarks across reasoning, mathematics, and truthfulness.

Language	Darija									English				
Model/Benchmark	Size (B)	DarijaMMLU	DarijaHellaSwag	Sentiment Analysis	GSM8K-darija-gen	Summarization				MMLU	TruthfulQA	HellaSwag	GSM8K @5	GSM8K-gen
	Metrics	Acc	Acc	Acc	Acc	chrF	ROUGE-1	ROUGE-L	BERTScore	Acc	bleu_acc	Acc	Acc	Acc
QWEN2.5-7B-INSTRUCT	7	44.9	38.5	63.6	43.9	26.5	9.4	9.1	36.7	68.7	63.1	65.4	75.8	90.1
QWEN2.5-7B-INSTRUCT-DARIJA	7	52.7	45.5	60.4	69.8	27.4	8.2	8	39	70	53.6	73.9	74.6	87.2
QWEN2.5-14B-INSTRUCT	14	57.5	45.9	63.3	72.3	26.3	9.1	8.8	36.2	76.8	70.8	75.6	80.2	94.2
QWEN2.5-14B-INSTRUCT-DARIJA	14	57.6	51.3	64.1	83.3	27.1	7.3	7.1	38.8	75.3	61.7	79.1	75.6	92.4
QWEN2.5-32B-INSTRUCT	32	63.9	45.9	65.8	78.1	27.2	9.4	9.2	37.2	73.9	70.5	74	78.6	90.5
QWEN2.5-32B-INSTRUCT-DARIJA	32	59.8	53.1	64.8	84.1	27.4	7.5	7.3	38.9	77.8	60.1	79.9	79.8	90

Table 2: QWEN-2.5 ablation

with distinct magnitudes and stability:

- **GEMMAROC-4B-LIMA** begins to activate minimal latent traces of Darija present in the pretrained backbone. However, the improvement remains limited (+2.1 pp on *DarijaMMLU*) and unstable, particularly in sentiment analysis (–16.8 pp), suggesting that such minimal exposure does not suffice for robust alignment.

- **GEMMAROC-4B-DEITA** shows a significant uplift across Darija benchmarks: +9.9 pp on *DarijaMMLU* and +8.0 pp on *DarijaHellaSwag*, while maintaining the original English reasoning performance (*MMLU* $\approx$ 51%). This indicates that the model successfully internalizes dialect-specific structures without catastrophic forgetting of its initial knowledge.

- **GEMMAROC-4B-TULU** further enhances both Darija and reasoning capabilities, reaching 47.5% on *DarijaMMLU* and 47.1% on *DarijaHellaSwag*. Compared to GEMMAROC-4B-DEITA, it brings an additional +4.9 pp on *DarijaMMLU* and +2.9 pp on *DarijaHellaSwag*, while also improving English reasoning (+2.8 pp on *GSM8K* and +5.0 pp on *HellaSwag*). The slight decline in sentiment accuracy suggests that the limited sentiment content in TULU may not fully offset the domain shift, but overall cross-lingual performance remains intact. This demonstrates that reasoning-oriented finetuning promotes deeper cross-lingual transfer while stabilizing bilingual competence.

Scaling Effects (27B models). At larger scale, the same pattern holds when moving from GEMMA3-27B-IT to GEMMAROC-27B. Finetuning on the Tulu corpus lifts performance from 55.65% to 61.61% on *DarijaMMLU* (+5.96 pp) and from 49.13% to 62.33% on *DarijaHellaSwag* (+13.2 pp), while maintaining strong results in English reasoning (*MMLU* = 73.6%, *GSM8K* = 84.2%). The improvement also extends to Darija reasoning, with *GSM8K-darija-gen* increasing from 82.71% to 84.46%. These findings confirm that Darija finetuning significantly enhances dialectal competence while preserving the pretrained backbone's general reasoning abilities, limiting the typical catastrophic forgetting observed in multilingual adaptation.

Cross-Model Comparison. A direct comparison with ATLAS-CHAT-27B further contextualizes these gains. Despite comparable model sizes, GEMMAROC-27B reaches similar or higher levels on key Darija reasoning benchmarks, achieving 61.61% on *DarijaMMLU* (vs. 61.95% for ATLAS-CHAT) and substantially outperforming it on *DarijaHellaSwag* (+13.96 pp) and *GSM8K-darija-gen* (+13.4 pp). While ATLAS-CHAT maintains higher sentiment accuracy (73% vs. 59.3%), GEMMAROC demonstrates superior reasoning coherence and bilingual stability, highlighting the benefit of controlled Darija-targeted finetuning over large-scale multilingual exposure.

Performance on sentiment and summarization tasks remains variable, reflecting the scarcity of native Darija data in these domains rather than a flaw of the alignment method. We deliberately scope this work to a reasoning-dominant evaluation, which we identify as the most transferable

Metric	ATLAS-CHAT-27B	GEMMAROC-27B
Helpfulness & profile alignment	3.17	**3.97**
Accuracy	3.67	**4.23**
Completeness	2.93	**3.90**
Darija fluency	3.93	**5.00**
Weighted composite (w above)	3.58	**4.35**

Table 3: Moroccan-law study (30 prompts, 10/profile). Scores are 1–5 weighted model means.

and measurable dialect competence

5.2 Qualitative Evaluation Results

Table 3 reports the Moroccan-law evaluation, where GEMMAROC-27B consistently surpasses ATLAS-CHAT-27B across all rubrics, with a +0.77 absolute gain on the 5-point composite scale (+21.5 %). The largest improvements occur in *Completeness* (+0.97; +33.1 %) and *Darija fluency* (+1.07; +27.2 %), showing that GEMMAROC delivers more detailed, natural citizen-facing responses.

GEMMAROC-27B produced fuller, better-structured answers, especially in marriage and labour-law cases, providing concrete steps, article references, and idiomatic Darija adapted to each user profile. ATLAS-CHAT remained terse or literal in MSA. Even when both were factually correct, GEMMAROC added procedural and contextual clarity, such as outlining remedies and protective measures in domestic-violence cases. These results confirm that reasoning-centric fine-tuning enables GEMMAROC-27B to extract, organize, and reframe dense legal text into accurate, complete, and natural guidance bridging formal law and everyday understanding.

5.3 Cost Efficiency

Fine-tuning the 27B-parameter version of GEMMA-3 required ≈48 GPU·h (8×H100 for 6 h), consuming ≈26 kWh and emitting 10 kg CO_2e (at 0.4 kg CO_2/kWh; Ember and Institute 2024). Monetary costs are omitted due to cloud-pricing variability, but compute remains two orders of magnitude below prior full-scale dialectal training. Training and evaluation losses decreased smoothly, with best validation at step 900 (∼2 epochs), confirming early convergence and minimal overfitting. Full energy calculations appear in Appendix A.3.

5.4 Backbone-Agnostic Ablation

To further validate fairness and generality, we extend the analysis to the QWEN-2.5 family, whose distinct tokenizer and training corpus make it an independent testbed for our method. We fine-tune QWEN2.5-7B, QWEN2.5-14B, and QWEN2.5-32B using the same setup: TULU (50K) translated into Darija with 30% English preserved, identical LoRA and hyperparameters as GEMMA. This isolates the recipe's effect from model-specific factors. Results in Table 2 show consistent gains across Darija and English benchmarks:

- **Darija gains.** All QWEN models improve with our recipe, except QWEN2.5-32B-INSTRUCT, whose *DarijaMMLU* slightly drops. QWEN2.5-7B improves from 44.9→52.7% on *DarijaMMLU* and 38.5→45.5% on *DarijaHellaSwag* with ≤6K instructions. Larger models reach 57.6% (14B) and 59.8% (32B), matching GEMMA.
- **Reasoning robustness.** QWEN2.5-14B/32B-DARIJA achieve 83.3% and 84.1% on *GSM8K-darija-gen*, confirming bilingual math transfer without forgetting.
- **English stability.** English benchmarks remain stable: *MMLU* 76.8% vs. 75.3% post-tuning; GSM8K and TruthfulQA vary within normal bounds.

These results confirm our recipe's backbone-agnostic nature, yielding strong Darija and bilingual reasoning even in 7B models with minimal supervision and scaling predictably with size.

6 Conclusion

We introduced a *reasoning-centric, data-efficient alignment* approach for adapting open LLMs to under-resourced dialects. We validated the approach with GEMMAROC, a family of Moroccan Arabic (Darija) models built on the GEMMA-3 backbone and benchmarked against ATLAS-CHAT. GEMMAROC-27B matches ATLAS-CHAT on *DarijaMMLU* (61.6%) and surpasses it on *HellaSwag* (62.3% vs. 48.4%), while showing only moderate trade-offs on English reasoning benchmarks. Training required only *48 GPU·hours*, using one-tenth the supervision and 50× lower energy consumption.

Beyond empirical gains, our results introduce a new perspective on dialect adaptation: reasoning-centric alignment acts as an effective bridge between global and local language capabilities. This conceptual contribution extends the scope of alignment research beyond high-resource multilingual settings to previously neglected dialects.

Limitations

Although the individual components of our recipe are known, their combined effect for dialectical reasoning alignment had not been empirically validated before this work. GEMMAROC excels in reasoning, matching or surpassing ATLAS-CHAT, and achieves strong, competitive results across most tasks. Some performance gaps remain in summarization, sentiment analysis, and reading comprehension, where richer Darija-specific data would be beneficial. The model was trained solely with parameter-efficient SFT, relies on machine-translated Darija samples with limited verification, and retains the original SentencePiece tokenizer, which is not optimized for Darija's script and linguistic nuances. A random 2 % subset was manually inspected by native speakers for semantic accuracy and script consistency. Future releases will include automatic quality metrics and public translation guidelines to support third-party auditing. aEvaluations focused on academic QA, common sense reasoning, and mathematics, without exploring long-context tasks or open-ended dialogue. Large-scale human evaluations in Darija are also pending, and the experiments were capped at 27 B parameters. Future work will explore full alignment pipelines, script- and language-aware tokenization, broader evaluations, and comprehensive human studies.

Ethical Considerations

Motivation and scope. This work aims to reduce the exclusion of under-resourced languages by eliciting robust Darija competence from open models with minimal data and energy. While the goal is inclusion and digital sovereignty, dialect-focused systems can also amplify harms if not designed and governed responsibly. Therefore, we discuss fairness, environmental impact, and transparency.

Bias, fairness, privacy, and provenance. Our Darija instruction data was translated with GEMINI, resulting in a relatively standardized variety of each language. This consistency is beneficial for training, but it underrepresents the regional Darija register dialects. While dialectal diversity is thus limited, the underlying instruction datasets (LIMA, DEITA, TULU) are highly curated, clean, and rich in reasoning and logic, especially TULU, helping to minimize harmful bias and ensuring strong alignment properties. Translations of LIMA, DEITA, and TULU follow the original datasets' open licenses (CC BY 4.0 / Apache 2.0), and no proprietary or personally identifiable material was included. Each translated sample retains metadata linking to its English source for traceability. Finally, we use only public, license-respecting sources, excluding any residual personal information.

Environmental impact. We prioritize low-compute alignment (LoRA, small high-quality datasets) and perform transparent energy accounting (GPU·h, estimated kWh and CO_2e), showing that competitive dialect performance can be achieved at orders-of-magnitude lower footprint than prior work. While estimates depend on datacenter and grid intensity, we report methodology, assumptions, and intermediate values (see Appendix A.3) to support verification and comparison. We encourage researchers to adopt similar accounting and to favor data-efficient recipes whenever feasible.

Transparency, and reproducibility. All components of this work are documented and reproducible. The full data-preparation pipeline, training configuration files, and evaluation scripts are archived and will be released under an anonymous repository link for review (and later mirrored publicly on Hugging Face after acceptance). Each artifact includes metadata on dataset provenance, translation prompts, and random-seed initialization. Package versions and environment details are listed in Appendix A to enable deterministic re-runs. The models are intended for educational, research, and public-service applications in Morocco and related dialect communities. They must not be deployed in surveillance, misinformation, or political-profiling contexts.

Use of generative assistants. We used generative language assistants, specifically GPT-5 and Gemini, to improve the clarity and fluency of the English writing, as we are not native English speakers. All ideas, analyses, and conclusions presented in the paper originate entirely from the authors; the models were employed solely for linguistic refinement and not for data generation, labeling, or result creation.

References

International Energy Agency. 2024. Electricity 2025 – Emissions Annex. Average global grid intensity 2024: 0.40 kg CO_2/kWh.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 2 others. 2025. Allam: Large language models for arabic and english. In *International Conference on Learning Representations*.

Abdelaziz Bounhar and Abdeljalil El Majjodi. 2025. Al-atlas: A causal language model for moroccan darija. https://huggingface.co/atlasia/Al-Atlas-0.5B/.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *Preprint*, arXiv:2401.07037.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Yicheng Chen, Yining Li, Kai Hu, Zerun Ma, Haochen Ye, and Kai Chen. 2025. Mig: Automatic data selection for instruction tuning by maximizing information gain in semantic space. *arXiv preprint arXiv:2504.13835*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. Describes the Vicuna chat model and its USER/ASSISTANT conversational format.

Karl Cobbe and et al. 2021. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*.

Team Cohere, :, Aakanksha, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, Zahara Aviv, Sammie Bae, Saurabh Baji, Alexandre Barbet, Max Bartolo, Björn Bebensee, and 211 others. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

NVIDIA Corporation. 2024. NVIDIA A100 Tensor Core GPU Datasheet. URL: https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf.

Devleena Das and Vivek Khetan. 2024. DEFT-UCS: Data efficient fine-tuning for pre-trained language models via unsupervised core-set selection for text-editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20296–20312, Miami, Florida, USA. Association for Computational Linguistics.

Databasemart. 2025. Optimizing vllm performance on A100 80 gb: Benchmark insights. https://www.databasemart.com/blog/vllm-gpu-benchmark-a100-80gb. Accessed May 19, 2025.

Ember and Energy Institute. 2024. Carbon intensity of electricity generation. https://ourworldindata.org/grapher/carbon-intensity-electricity. Accessed May 19, 2025.

Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. 2025. Automatic Instruction Data Selection for Large Language Models via Uncertainty-Aware Influence Maximization. In *Proceedings of the ACM Web Conference 2025*, pages 4969–4979, Sydney, Australia. ACM.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *CoRR*, abs/2002.05651. Revised version published 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2024. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Preprint*, arXiv:2409.12958.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency

in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 546–566, Bangkok, Thailand. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

LMSYS Org. 2023. Fastchat: An open platform for training, serving, and evaluating large language models. Implements the Vicuna prompt/chat template used for multi-turn formatting.

Yuren Mao, Yaobo Liang, Nan Duan, Haobo Wang, Kai Wang, Lu Chen, and Yunjun Gao. 2022. Less-forgetting multi-lingual fine-tuning. In *Advances in Neural Information Processing Systems 35: NeurIPS 2022, New Orleans, USA*.

MDTAR. 2024. Ministry of Digital Transition and Administration Reform, Digital Morocco 2030: National Strategy. https://www.mmsp.gov.ma/sites/default/files/2024-09/PlaquetteInstitutionnel_18092024_Ang.pdf. Accessed: 2025-05-13.

NVIDIA. 2024. NVIDIA A100 Tensor Core GPU datasheet. https://www.nvidia.com/content/dam/.../a100-80gb-datasheet.pdf. Accessed May 19, 2025.

ProtectAI. 2025. protectai/xlm-roberta-base-language-detection-onnx. https://huggingface.co/protectai/xlm-roberta-base-language-detection-onnx. Accessed May 2025. ONNX version of papluca/xlm-roberta-base-language-detection using the Optimum library. Fine-tuned on the Language Identification dataset.

Lu Qin and 1 others. 2025. A survey of multilingual large language models. *PeerJ Computer Science (via PMC)*, 11:e11783891.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource moroccan arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20829–20855, Miami, Florida, USA. Association for Computational Linguistics.

Junhong Wu, Yang Zhao, Yangyifan Xu, Bing Liu, and Chengqing Zong. 2025. Boosting LLM translation skills without general ability loss via rationale distillation. In *Submitted to ACL Rolling Review - February 2025*. Under review.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, and Gholamreza Haffari. 2024. MIXTURE-OF-SKILLS: Learning to optimize data usage for fine-tuning large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14226–14240. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of ACL*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS-2023)*, volume 36, pages 55006–55021.

A Appendix

A.1 SFT Dataset Preparation

A.1.1 Translation Prompt Template

To ensure consistency and high quality in automatic translation from English to Moroccan Darija, we employ a strict prompt with clear translation and preservation guidelines. The full prompt, as used for dataset generation, is presented below.

TRANSLATION_PROMPT =
Translate the following text into Moroccan Darija, written in Arabic script, following these strict instructions:
DO NOT TRANSLATE the following elements (keep them exactly as they are):

1. **Code blocks**:
   Keep *all programming code* (Python, JS, etc.) **unchanged**.
   Translate comments inside code blocks to Darija (Arabic script).

2. **Math formulas and LaTeX**:
   Preserve all mathematical notation and LaTeX equations. Do not alter them in any way.

3. **Technical Terms**:
   Keep technical/scientific terms in English unless they have a known Arabic equivalent.
   For example:

   - Use "

     صيغة" for "formula"
   - Use "

     معادلة" for "equation"
   - Use "

     دالة" for "function"
   - But do not translate terms like "API", "eigenvalue", "html", etc.

4. **Proper Names**:
   Do not translate names of people, tools, or software (e.g., Python, Newton, GitHub, etc.).

TRANSLATE into Moroccan Darija (Arabic script):

- Regular sentences, explanations, and natural narrative text.
- Comments inside code blocks (e.g., `# comment here`).
- Use a natural mix of Moroccan Darija and Modern Standard Arabic where appropriate:
  - Use Darija for informal or conversational parts.
  - Use Standard Arabic for formal or widely recognized scientific terms. Do your best to find the most suitable translations.
  - If translating certain phrases would sound awkward, unclear, or artificial in Darija, then translate them into Standard Arabic.

GOAL:
Make the content clear and naturally understandable to Moroccan Arabic speakers while respecting the technical integrity of the original text. Return the translated text without saying anything else.
Examples:
Example 1 (Code + Explanation)
Input:

The 'Function' calculate_mean(data) is used to compute the average of a list of numbers.

```
// This function iterates through the list
// and calculates the sum. Python

def calculate_mean(data_list):
    # Calculate sum of elements
    total_sum = sum(data_list)
    # Calculate number of elements
    count = len(data_list)
    # Return the mean
    return total_sum / count
```

Its 'Complexity' is O(n).

Output: ال'Function' calculate_mean(data)
كتستعمل باش نحسبو المعدل
ديال شي ليستة ديال الأرقام.
هاد الدالة كدور على العناصر //
وكتحسب المجموع. //
Python

```
def calculate_mean(data_list):
    # Calculate sum of elements
    total_sum = sum(data_list)
    # Calculate number of elements
    count = len(data_list)
    # Return the mean
    return total_sum / count
```

ال'Complexity' ديالها هي O(n).

Example 2 (Math Formula)
Input:

Consider the quadratic equation $ax^2 + bx + c = 0$. The solutions can be found using the formula:
$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. This is known as the quadratic formula.

Output:

ناخدو المعادلة من الدرجة الثانية
$ax^2 + bx + c = 0$.
الحلول ديالها كتلقاو بهاد الصيغة:
$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.
هادي معروفة بصيغة المعادلة من الدرجة الثانية.

Example 3 (Fully Regular Text)
Input:

This chapter introduces the basic concepts. Please read it carefully.

Output:

هاد الجزء كيعرف بالمفاهيم الأساسية.
عافاك قراه مزيان.

Now, translate the following text:

A.1.2 Comprehensive Description of Source Datasets

Comprehensive Description of LIMA-1k Dataset:

The LIMA-1k dataset, introduced in (Zhou et al., 2023), comprises 1,000 meticulously curated examples specifically designed to efficiently align large language models (LLMs) with minimal computational resources. This dataset originates from an initiative exploring the hypothesis that effective alignment and instruction-tuning do not necessarily require extensive datasets or reinforcement learning from human feedback (RLHF). Instead, the primary goal of LIMA-1k is to demonstrate that alignment can predominantly be achieved by leveraging the knowledge already embedded within pre-trained models, with fine-tuning merely instructing the model in stylistic and structural nuances for interacting effectively with users.

LIMA-1k is curated from several high-quality online sources, notably Stack Exchange, wikiHow, and selected Reddit communities (r/WritingPrompts), augmented with manually authored examples to enhance stylistic uniformity and thematic diversity. Table 4 summarizes the dataset's key statistics.
Below are representative multi-turn dialogue examples illustrating the conversational coherence and stylistic uniformity achieved by LIMA-1k.
Example 1:
Human: Tell me an interesting fact about geography.

Assistant: Here are some interesting facts about geography:

- **Bhutan** is one of the most intriguing places in the world. The Buddhist kingdom is the only carbon-negative country on Earth, with strict environmental laws and cultural norms that preserve its stunning natural landscape. For similar reasons, Bhutan was one of the last countries to introduce television in 1999, and has no traffic lights to this day.

- There are currently few territories on Earth that are unclaimed by any sovereign state (*terra nullius*). One example is **Bir Tawil**, along the border between Egypt and Sudan, which was created by a discrepancy between borders drawn in 1899 and 1902. The territory remains unclaimed by either country because it would entail forfeiting the much larger Halaib Triangle, which is claimed by both countries. Bir Tawil has no settled population, but the land is used by Bedouins who roam the area.

- The **antipode** of any spot on Earth is the point directly opposite to it on the globe. Interestingly, most land areas' antipodes fall in the ocean. For instance, the antipode of Spain is in New Zealand, but the antipode of most locations in North America lies in the Pacific Ocean.

Example 2:
Human: I am planning to start a book club with

Data Source	Examples	Avg. Input Length	Avg. Output Length
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift Reddit r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Manually Authored (Group A)	200	40	334
Total/Average	**1,000**	**93**	**727**

Table 4: Summary statistics of the LIMA-1k dataset (Zhou et al., 2023).

some friends. Can you write an email invitation for the same?

Assistant: Here is an example email invitation for a book club:

> **Subject:** Join me for a book club!
>
> Hi [Friend's Name], how are you?
>
> I'm starting a book club with some friends, and I'd love to have you join us!
>
> The book club is a casual and fun way to read and discuss books that we love or want to learn more about, and also make new friends. We plan to meet more-or-less once a month.
>
> We're thinking of having our first meeting on **[Date & Time]**. Could that work for you?
>
> Would love to see you there!
>
> Thanks,
> **[Your Name]**

The significance of LIMA-1k lies in its exceptional efficiency and effectiveness. Remarkably, despite its modest size of only 750,000 tokens, models fine-tuned on LIMA-1k exhibit robust generalization capabilities and compete closely with extensively trained RLHF models. This dataset convincingly demonstrates the practicality of minimal yet highly curated datasets for achieving high-quality alignment, substantially reducing computational costs and democratizing the capability for effective model alignment.

Comprehensive Description of DEITA-6k Dataset

DEITA-6k is a meticulously curated dataset of *6,000 instruction-response pairs* selected from a pool of *300,000 high-quality public instruction datasets*, including ShareGPT (58k), UltraChat (105k), and WizardLM (143k). It is designed to be the *smallest effective collection* that still achieves state-of-the-art alignment quality after SFT on 7B- and 13B-parameter models.

Selection Methodology The DEITA pipeline employs three criteria for selection:

- **Complexity (c):** Measured using an EVOL-COMPLEXITY scorer trained to evaluate the reasoning depth of instructions.
- **Quality (q):** Assessed using an EVOL-QUALITY scorer focusing on the helpfulness and informativeness of responses.
- **Diversity:** Ensured by applying a cosine similarity threshold ($\tau = 0.9$) over sentence embeddings to avoid redundancy.

The final score $s = c \times q$ ranks the samples, and the highest-ranked examples satisfying diversity constraints form the DEITA-6k dataset.

Key Statistics:

Statistic	Value
Total Examples	6,000
Source Pools	ShareGPT (58k), UltraChat (105k), WizardLM (143k)
Average Input Length Ceiling	2,048 tokens
Domain Coverage	Balanced across 8 MT-Bench categories
Complexity Gain vs Random	+0.43 MT-Bench (6.46 → 6.03 on LLaMA-13B)
Quality Gain vs Random	+3.17% AlpacaEval (77.08% → 73.91%)
Scaling Insight	3K DEITA items match the performance of 300K examples

Table 5: Key statistics of the DEITA-6k dataset. Complexity and quality gains are measured relative to a random 6k subset of the same source pool.

Representative Examples:

Example 1: Entain's Competitors Discussion

Human: list me some of Entain's competitors in their industry
Assistant: Entain, formerly known as GVC Holdings, is a multinational sports betting and gaming company. Competitors include:
1. Flutter Entertainment
2. William Hill
3. Bet365

Human: competitors who are also expanding into esports, game development
Assistant: Companies include:
1. DraftKings
2. FanDuel
3. Betway
4. Unikrn

Human: what studios has Unikrn acquired? Furthermore, elaborate on äcquired.
Assistant: As of my knowledge cutoff, Unikrn has not acquired any gaming studios. Explanation of acquisitions:
1. Tencent/Riot Games: Full acquisition over time.
2. Tencent/Supercell: Majority stake through a consortium.
3. Activision Blizzard: Complete acquisition of Vivendi Games.

Example 2: IP Ownership and Business Model

Human: I have ownership of the IP of video games created by S2 Games, excluding Heroes of Newerth... What is the appeal of this, to any investor?
Assistant: Key appeals include:
1. Nostalgia and a loyal player base.
2. Unique IP and lore for expansion.
3. Use of AI-generated art and modern features.
4. Platform potential for indie developers.
5. Revenue-sharing model incentives.
6. Strong growth potential with community engagement.

Example 3: Data Warehouse Simplifications

Human: explain data warehouse architecture in easy language
Assistant: A data warehouse collects data from various sources, transforms it, and stores it centrally. Main components:
1. Data Sources
2. ETL Process
3. Data Warehouse
4. OLAP System
5. Front-End Tools

Example 4: Contracted Chief Brand Officer's Role

Human: Generate a description of a Contracted Chief Brand Officer's role
Assistant: A Contracted Chief Brand Officer becomes an integral part of your executive team, leading brand strategy and development, working with internal teams and external creatives to build a competitive brand presence, and providing ongoing strategic advisement.

Summary:
DEITA-6k demonstrates that *quality, complexity, and diversity can be engineered simultaneously* to produce a highly effective dataset for alignment. Despite its compact size, it enables language models to achieve near-SOTA performance with significantly reduced computational costs.

Comprehensive Description of TULU-50k Dataset

Comprehensive Description of TULU–50K Dataset

Origin and Purpose:
TULU–50K is a 50 000-example SFT subset distilled from the public TÜLU 3 SFT pool (≈939 k dialogues) released by the Allen Institute for AI. The subset was produced with the *MIG* (Maximising Information Gain) sampler (Chen et al., 2025), which automatically balances *quality*, measured by DEITA scores, and *diversity* in a semantic label-graph. Its primary goal is to offer an open, compact yet high-coverage instruction-following corpus for research on post-training large language models, especially where full-scale SFT mixes are computationally prohibitive. The data are licensed ODC-BY and are entirely in English.
Selection Method:
MIG first tags every dialogue in the 939 k-sample pool with semantic topic labels, then scores each instance with DEITA quality metrics. It models the

tag space as a weighted graph, propagates information across semantically-related tags, and greedily selects examples that yield the highest marginal information gain, provably achieving a $(1 - 1/e)$ approximation of the optimal subset. Compared with random or purely quality-based filtering, the resulting 50 k set delivers state-of-the-art downstream scores while cutting sampling time by two orders of magnitude.

Some Examples:

Example 1:

Human: Captain Jake, a somewhat short-tempered commercial airline pilot, has to navigate a complex flight path to avoid airspace controlled by airlines embroiled in corporate squabbles. The flight path consists of a series of vectors in 3-dimensional space. Each vector represents a leg of the journey from one waypoint to another.

- **1.** Given the following vectors representing legs of the journey:
  - Vector A: $\mathbf{a} = (3, 2, -1)$
  - Vector B: $\mathbf{b} = (-2, 4, 5)$
  - Vector C: $\mathbf{c} = (1, -3, 2)$

  Calculate the total displacement vector $\mathbf{d}$ from the start to the end of the journey. Express $\mathbf{d}$ in its simplest form.

- **2.** Captain Jake needs to maintain a steady altitude change to avoid turbulence caused by corporate squabbles in certain airspace layers. If the journey described by vector $\mathbf{d}$ must ensure that the vertical displacement (z-component of $\mathbf{d}$) is zero, determine the required adjustment vector $\mathbf{e} = (x, y, z)$ such that the modified displacement vector $\mathbf{d}' = \mathbf{d} + \mathbf{e}$ satisfies this condition. Additionally, compute the magnitude of the adjustment vector $\mathbf{e}$.

Assistant: Let's break down the problem into the given tasks and solve each part step by step.

1. **Calculate the total displacement vector $\mathbf{d}$:**

   Given vectors:

   $$\mathbf{a} = (3, 2, -1), \quad \mathbf{b} = (-2, 4, 5),$$
   $$\mathbf{c} = (1, -3, 2)$$

   The total displacement vector is:

   $$\mathbf{d} = \mathbf{a} + \mathbf{b} + \mathbf{c}$$

   Compute each component:

   $$\mathbf{d}_x = 3 + (-2) + 1 = 2,$$
   $$\mathbf{d}_y = 2 + 4 + (-3) = 3,$$
   $$\mathbf{d}_z = -1 + 5 + 2 = 6$$

   So, the total displacement vector is:

   $$\mathbf{d} = (2, 3, 6)$$

2. **Determine the adjustment vector $\mathbf{e}$:**

   We want $\mathbf{d}' = \mathbf{d} + \mathbf{e}$ such that the z-component of $\mathbf{d}'$ is zero:

   $$6 + z = 0 \implies z = -6$$

   Thus, $\mathbf{e} = (x, y, -6)$. Since only the vertical displacement needs correction, we can choose $x = 0$ and $y = 0$ for simplicity. Therefore:

   $$\mathbf{e} = (0, 0, -6)$$

   The magnitude of $\mathbf{e}$ is:

   $$|\mathbf{e}| = \sqrt{0^2 + 0^2 + (-6)^2} = \sqrt{36} = 6$$

   Final Answer: The displacement vector is $\mathbf{d} = (2, 3, 6)$, and the magnitude of the adjustment vector $\mathbf{e}$ is 6.

Example 2:

Human: A dedicated spectator named Alex spends their weekends analyzing basketball player statistics to provide data-driven recommendations. Alex has been tracking the performance of a team throughout the season and has compiled a dataset consisting of the following metrics for each player: points per game (PPG), assists per game (APG), rebounds per game (RPG), and a newly devised metric called *Efficiency Rating* (ER), which is calculated as:

$$\text{ER} = (PPG \times 0.4 + APG \times 0.3 + RPG \times 0.3)$$

- **1.** Alex wants to determine the optimal combination of players in a 5-person lineup to maximize the team's average Efficiency Rating. Given that there are 10 players on the team, develop an algorithm or mathematical model that can identify the best combination. What is the maximum possible average Efficiency Rating for the team?

- **2.** After selecting the optimal lineup, Alex decides to project the team's overall performance in the upcoming playoffs. Alex uses a regression model where the team's winning probability in a game is given by:

  $$P(\text{win}) = 0.5 + 0.1 \times (\text{average ER of the lineup})$$

  Assuming the optimal lineup maximizes the team's average ER to 25, calculate the probability of the team winning a best-of-7 series, where the team must win 4 out of 7 games to advance.

Assistant: Let's solve the problem in two parts.

1. **Optimal Lineup to Maximize Average ER:**
   - Sort the 10 players by their Efficiency Rating (ER).
   - Select the top 5 players with the highest ER values:

     $$ER_1 \geq ER_2 \geq ER_3 \geq ER_4 \geq ER_5$$

   - Calculate the maximum average ER:

     $$\frac{ER_1 + ER_2 + ER_3 + ER_4 + ER_5}{5}$$

   - According to the problem, the maximum average ER is given as 25.
2. **Projecting the Team's Winning Probability:**

   Using the regression model:

   $$P(\text{win}) = 0.5 + 0.1 \times 25 = 0.5 + 2.5 = 3.0$$

   Since probabilities cannot exceed 1, the model saturates at:

   $$P(\text{win}) = 1$$

   Probability of Winning a Best-of-7 Series:

   - The team must win 4 out of 7 games. - If $p = 1$, the team is certain to win each game. - Therefore, the probability of winning the series is effectively:

   $$P(\text{series win}) = 1$$

   Final Answer: The maximum possible average Efficiency Rating is 25, and the team has a theoretical 100% chance of winning the best-of-7 series under the given model.

Concluding Remarks:
TULU–50K strikes a pragmatic balance between dataset footprint and skill coverage. Empirically, models fine-tuned on this 50K slice match or exceed full-pool baselines on both knowledge-centric (ARC, MMLU) and preference-centric (AlpacaEval v2, MT-Bench) benchmarks while slashing compute and curation time. Its transparent provenance, permissive license, and rich per-sample metadata make it an efficient, community-friendly test-bed for exploring data-centric alignment questions and rapid prototyping of instruction-tuned LLMs.

A.2 Illustrative Examples from Final Darija Datasets

A.2.1 Summary Statistics of Our SFT Datasets

Dataset	Total Samples	Darija Translated	Original English
LIMA 1K	1,000	700 (70%)	300 (30%)
DEITA 6K	5,000	3,700 (74%)	1,300 (26%)
TULU 50K	46,000	33,000 (72%)	13,000 (28%)

Table 6: Dataset sizes after filtering/translation.

A.2.2 Illustrative Examples from Darija Lima

See Figures 3 and 4.

A.2.3 Illustrative Examples from Darija Deita

See Figures 5, 6, and 7.

A.2.4 Illustrative Examples from the Darija TULU

See Figures 8 and 9.

A.3 Efficiency and carbon footprint

Protocol. We follow the GREEN 500 methodology and the recommendations of Strubell et al. (2019); Henderson et al. (2020). The total electrical energy E for a training run is

$$E = \sum_{g \in \{\text{A100,H100}\}} N_g \, t_g \, P_g \, \text{PUE} \, \eta,$$

where N_g and t_g are the number of GPUs and wall–clock hours, P_g is the average power draw (350 W for A100-80 GB (Corporation, 2024), 500 W for H100-80 GB), PUE $= 1.3$ is a modern data center power-usage effectiveness, and $\eta = 0.9$ down-scales for the fact that accelerators seldom sustain their full TDP for an entire job.

Our training budget. Table 7 lists every experiment reported in Section 4. All ablations on the 4 B backbone consume a *combined* 10 GPU·h on A100-80 GB, while the GEMMAROC run uses 48 GPU·h on H100-80 GB. Plugging these figures into the formula above gives:

$$E_{\text{A100}} = 10\,\text{h} \times 0.35\,\text{kW} \times 1.3 \times 0.9 \;=\; 4.1\,\text{kWh},$$
$$E_{\text{H100}} = 48\,\text{h} \times 0.50\,\text{kW} \times 1.3 \times 0.9 \;=\; 28.1\,\text{kWh},$$
$$E_{\text{total}} \approx 32\,\text{kWh}.$$

With a 2024 global mean grid-carbon intensity of 0.40 kg CO_2/kWh (Agency, 2024) this implies

$$C_{\text{ours}} \;\approx\; 13\,\text{kg CO}_2\text{e}.$$

For clarity, the GEMMAROC trained on the 8 H100 GPUs alone amounts to 48 GPU·h, corresponding to roughly 26 kWh of electricity consumption and about 10 kg of CO_2e.

Baselines: ATLAS-CHAT. ATLAS-CHAT-27B used LoRA rank 256, 3 epochs over $500\,\text{k}$ sequences of length 2 048 ($\simeq 675\,\text{M}$ training tokens), and used the GEMMA-2-27B as its base model. A recent vLLM benchmark on a single A100-80 GB reports an *inference* throughput of $\approx 5{\times}10^2$ tok s^{-1} for GEMMA-2-27B (Databasemart, 2025). Training carries out a forward *and* a backward pass; empirical FLOP analyses therefore divide inference speed by $\sim$3. We conservatively adopt $r = 55$ tok s^{-1} GPU^{-1}.

$$t_{\text{wall}} = \frac{675 \times 10^6}{8\,r\,3600} \;\approx\; 4.26 \times 10^2\ \text{h},$$
$$t_{\text{GPU}} = 8\,t_{\text{wall}} \;\approx\; 3.41 \times 10^3\ \text{GPU-h}.$$

Using the NVIDIA A100-80 GB SXM4 average training draw $P = 0.35\,\text{kW}$ (NVIDIA, 2024) and an industry-average PUE of 1.3 with 90 % utilization, the energy and carbon footprints are (Ember and Institute, 2024):

$$E = t_{\text{GPU}}\,P\,\text{PUE}\,0.9 \;\approx\; 1.4\ \text{MWh},$$
$$C = E \times 0.436\ \text{kg CO}_2\text{e kWh}^{-1} \;\approx\; 610\ \text{kg CO}_2\text{e}.$$

Finally, at a public on-demand rate of \$2 A100-h^{-1} the run would cost $\sim$ \$6.8k, underscoring the budget advantage of our own (far cheaper) fine-tune.

A.4 Training and Eval Loss of GEMMAROC-27B

Run	GPUs	Hours	GPU·h	Purpose
GEMMA-3-4B–LIMA	2×A100	0.4	0.8	1 K single-turn
GEMMA-3-4B–DEITA	2×A100	0.8	1.6	6 K dialogue
GEMMA-3-4B–DEITA^{+}	2×A100	0.8	1.6	+30
GEMMA-3-4B–TULU	2×A100	3	6	50 K reasoning
Subtotal 4 B			**10**	
GEMMA-3-27B–TULU	8×H100	6	48	final GEMMAROC
Total	—	—	**58**	

Table 7: Revised energy-accounting input reflecting the updated 10 A100 GPU·h for all 4 B experiments.

كيفاش تكون سعيد فزواج ماشي سعيد؟

الناس كيبقاو فزواج ماشي سعيد لأسباب بزاف. إذا كنت واحد من هاد الناس، تقدر تحس براسك ما يمكنش تكون فرحان. على الرغم من هادشي، تقدر تلقى طريقتك الخاصة للسعادة، حتى فشي وضعية خايبة، بممارسة عادات اللي كتوصل للسعادة، وحتى تقدر تخدم على الزواج ديالك باش توصل للسعادة كزوجين.

##تعلم تكون سعيد

.1قلب على طرق باش تكون ممتن. ماشي ديما ساهل تكون ممتن، خاصة وسط علاقة خايبة. على الرغم من هادشي، الامتنان يقدر يعاونك باش تواجه العلاقة الخايبة، ويقدر يوصلك باش تكون سعيد. خود الوقت كل نهار باش تعرف شنو ممتن ليه. حاول تكتب شي حاجات ممتن ليهم كل نهار فواحد الجورنال. تقدر تستعمل حتى المنشورات اللي فال social media لهاد الغرض. غالبا، واخا تكون الوضعية خايبة، تقدر تلقى شي حاجة تكون ممتن ليها. مثلا، تقدر مايعجبكش كيفاش كيتعاملو معاك فالعلاقة اللي فيها، ولكن تقدر تكون ممتن لأن عندك استقرار مالي فحياتك ف هاد الوقت. أولا، تقدر تكون ممتن لأن الطرف الآخر فالعلاقة كيبقى أب أو أم مزيانة لولادك.

.2شارك ف 'flow experiences'. لحظات ال Flow هي فاش كتنسى راسك فواحد التجربة بكونك مندمج كليا فشنو ما كنت كدير. إلا كنت فنان، كاتب، ولا حتى عداء، تقدر تكون ديجا فهمتي هاد النوع ديال التجارب. هاديك اللحظة اللي كيطيح فيها العالم، و أنت غير كتجرب و لا كتستمتع باللي كدير. الدراسات بينات باللي كلما كانو عندك لحظات Flow بزاف، كلما كنت سعيد بصفة عامة. اختار شي نشاط اللي كيتحدالك شوية، ولكن اللي يكون مألوف باش تقدر تنسى راسك فيه. مثلا، إلا كنت كتستمتع برسم المناظر الطبيعية، تقدر تجرب ترسم موضوع جديد بحال شي بورتريه ولا شي طبسيل ديال الفواكه.

.3حبس من نفس المشاكل. يعني، إلا كنت ديما كتلقى راسك كتجادل على نفس الحوايج، يمكن وصل الوقت باش تحط هاد الموضوع على جنب. خاصك تقرر باللي ماغاديش تناقش فيه حيت مايمكنش تتفقو و لا تحاولو تلقاو شي حل وسط اللي يخدم ليكم بجوج. مثلا، إلا كنتو كتجادلو بزاف على السياسة، تقدر تبغيو تخليو السياسة موضوع ممنوع. و لا، إلا كنتو كتجادلو بزاف على شنو من فيلم غادي تشوفو فالليلة ديال الجمعة، تقدر تبغيو تناوبو فالإختيار ديال الفيلم.

.4طور الإهتمامات الخاصة ديالك. إلا الزواج ديالك ماشي هو اللي بغيتي، يقدر وصل الوقت باش تلقى شي إشباع خارج الزواج ديالك وماشي على شكل خيانة. أن تكون عندك هوايات و اهتمامات خاصة غادي يعاونك باش تبقى مستقل وتبقى سعيد و مندمج مع العالم. فالحقيقة، تطوير الاهتمامات الخاصة فكرة مزيانة حتى إلا كان عندك زواج زوين. فكر أنك تستكشف اهتماماتك فالمكتبة، و لا تنضم لأندية الهوايات المحلية أو دروس الطبخ، و لا تاخد دروس مهنية فشي كلية مجتمعية قريبة .[...]

##خدم باش يكون عندك زواج أحسن

.1خصصو وقت باش تكونو مجموعين. تخصيص الوقت لبعضكم البعض طريقة باش تقولو أنكم ملتزمين ببعضكم البعض. قولو بأنكم غادي تقضيو وقت معين كل أسبوع مع بعضكم البعض، و بقاو عليه. غادي يعطيكم حتى فرصة باش تعرفو بعضكم البعض من جديد.

.2تفكرو شنو كنتو كتقدروا. فاش يالاه تلاقيتو، غالبا كنتو منجذبين جزئيا للإختلافات اللي كان كيعرضها الطرف الآخر. مثلا، تقدر تكون كتفرحك حقيقة أنه كان ولا كانت مندفعين و كيعجبهم يكونو عفويين. دابا، تقدر تلقى راسك كتكره هاد الصفة. المهم هو تحاول تتفكر علاش كنتي كتحب هاد الصفة فالأول و تولي تستمتع بيها من جديد. مثلا، تقدر تطلع ليك فالراس فاش الزوج ديالك يبغي يخلي كلشي و يسافر للجبال. من جهة أخرى، كتخلي حياتك ما توليش مملة بزاف. حاولو تلقاو توازن، و استمتعو باللي قدرتوا عليه .[...]

Figure 3: Illustrative example 1 from Darija Lima dataset.

شرح مبسط لencryption backdoors

كاينين جوج ديال المفاهيم لي شوية مستقلين على بعضياتهم ف backdooring ديال:encryption algorithms

*ال backdoor يقدر يكون explicit ولا .implicit ال explicit backdoor هو واحد لي كلشي عارف باللي راه كاين. أما ال implicit backdoorكيحاول يبقى مخفي على مالين ال .algorithm بطبيعة الحال، فاش كيكون شيexplicit backdoor ، الناس كيميلو يتجنبو ال algorithm بصفة عامة، يعني explicit backdoors ممكن يكون عندهم أمل باش يخدمو غير فوجود شي إطار قانوني لي كيفرض على implementers يستعملو ال system لي فيه ال .backdoor مثال على شي explicit backdoorهو ال Clipper chip (لي تخلى عليه ف الأخير). ال backdoor ماشي بالضرورة يكون ف الalgorithm ، كثر ف التجميع ديال algorithms ف شيprotocol ، و تقنيا كان عبارة على .automatic key escrowing method أما بالنسبة ل شيimplicit backdoor ، شوف ال Dual EC DRBG كمثال مشهور: كان كيخدم غير إلى مكان تا واحد عارف باللي فيه backdoor.

*ال security ديال ال backdoor يقدر يكونquantifiable ، ولا لا. ف حالة ديالDual EC DRBG ، ال mechanism كيستعمل طرق رياضية معروفة: ال NSA كانت عارفة باللي استغلال ال backdoor كيتطلب معرفة ديال شي internal secret key، مبني على) discrete logarithm على.(elliptic curves

شي security ماشي quantifiable هو لي كيوقع فاش كتحاول تدفع شي algorithm لي فيه عيب متعمد، ولا شي واحد لي عارف عليه شي cryptanalytic method لي ما نشرتيهش. هادي لعبة خطيرة بزاف بالنسبة ل شي وكالة تجسس، حيت مايمكنش تعرف واش أطراف آخرين يقدروا يلقاو العيب ولا لا. بحال هاد ال backdooring كيقلب عليك ف المدى الطويل.

لي مثير للاهتمام، ال NSA ما كتميلش تستعمل .non-quantifiable backdoors مثال مزيان هو .DES ف الوقت لي تصمم فيه، ال NSA كانت معتقدة باللي تقدر تواجه شيupfront 256 exhaustive search ، و باللي حتى واحد آخر (خصوصا السوفيات) ماعندوش ال technology و ال budget باش يدير هادشي. ال NSA كانت عارفة حتى على شي cryptanalytic method جديدة .(differential cryptanalysis) يعني فاش ال NSA تدخلات ف التصميم ديالDES ، أصرت باش تنقص ال key من 64 ل 56) bits هادي إضافة ديال شيquantifiable backdoor ، و كانت واضحة بزاف، يعني شبه(explicit ، و أيضا باش تقوي التصميم ضد .differential cryptanalysis هادا مثال مزيان على كيفاش ال NSA تجنبات بوعي شي non-quantifiable backdoor.كاين حاجة وحدة لي وكالات التجسس كيخافو منها كثر من أنهم ما يقدروش يتجسسو: هي الفكرة باللي وكالات تجسس منافسة أخرى ممكن حتى هي تتجسس.

إذن شي backdoor حقيقي و مزيان هو لي كيستعمل maths باش يعطي شي quantifiable security ضد الاستعمال الغير مصرح به ديال ال .backdoor صعيب بزاف يكون عندك quantifiable security بلا ما تجعل ال .backdoor explicit "الأفضل ف فئتها" ف هاد المجال هيDual EC DRBG ، و حتى قبل الSnowden business ، ال cryptographers كانو كيشوفوها غريبة بزاف و شكاو بقوة ف شي تلاعب) شوف ال analyses ف 2007، 6 سنين قبل.(Snowden

Figure 4: Illustrative example 2 from Darija Lima dataset.

Human:

أهلا، بغيتك تشرح ليا ال certificate pinning process خطوة بخطوة بمنظور عام.

Assistant

أكيد! ها هي خطوة بخطوة شرح لعملية certificate pinning بمستوى عالي:

1. ال client (بحال web browser أو mobile app) كيبدا اتصال ب server (بحال website أو API).

2. ال server كيجاوب ب SSL/TLS certificate ديالو، اللي كيتضمن public key ومعلومات تعريفية أخرى.

3. ال client كيدير فحص لل certificate مع مجموعة ديال trusted root authorities (بحال ليستة ديال trusted Certificate Authorities مخزنة عند ال client).

4. إلا كان ال certificate موثوق، ال client كيكمل عملية SSL/TLS handshake وكيدير اتصال آمن مع ال server.

5. ال client كيفيريفي خلال هاد العملية، ال certificate ديال ال server مع مجموعة ديال pins معرفة مسبقاً، واللي متخزنة localement عند ال client.

6. هاد ال pins يقدروا يكونوا:

a) Public key pins: عبارة عن hash ديال ال public key ديال ال server (بحال SHA-256 hash)

b) Certificate pins: عبارة عن hash ديال ال certificate (بحال كامل SHA-256 hash)

7. إلا كان ال certificate ديال ال server كيطابق مع شي pin من ال pins اللي معرفين مسبقاً، الإتصال كيدار. و إلا ماكانش، ال client كيرفض الإتصال وكيسالي ال SSL/TLS handshake.

8. إلا ال client واجه فشل في pin validation ، يقدر يعطي للمستخدم رسالة تحذيرية أو ياخد إجراءات أخرى باش ينبه المستخدم للخطر الأمني المحتمل.

بإختصار، certificate pinningهي آلية كتسمح لل clients يتأكدو باللي كيتواصلو مع ال server المقصود ماشي محتال. عن طريق مقارنة SSL/TLS certificate ديال شي server مع مجموعة ديال pins معرفة مسبقاً، ال clients يقدروا يخففو من الخطر ديال man-in-the-middle (MITM) attacks وأنواع أخرى ديال certificate-based attacks.

Figure 5: Round 1 of an Illustrative example from Darija Deita dataset.

Human:

واش هادشي كيعني باللي أي API call كنستعملو فيه axios مثلاً غادي يكونsecure ؟

Assistant

شهادة التثبيت بوحدها ما كتضمنش بأن كل API call كتستعمل مكتبة بحال Axios غادي تكون آمنة، حيت كاينين ثغرات محتملة أخرى في التطبيق و الشبكة اللي ممكن تستغل. ولكن، certificate pinningتقدر تنقص بشكل كبير من خطر هجمات الوسيط (man-in-the-middle attacks) (MITM)، اللي هي واحد من طرق الهجوم الشائعة في بزاف ديال الاختراقات الأمنية.

ملي كطبق certificate pinning مع Axios ولا أي مكتبة HTTP client أخرى، مهم تبع أحسن الممارسات وتأكد بأن ال client مكونفيگوري باش يفيريفي ويثبت الشهادة ديال السيرفر بشكل صحيح. بالزيادة على هادشي، تدابير أمنية أخرى خاصها تطبق، بحال input validation وoutput encoding ، باش تنقص بشكل أكبر من خطر الثغرات الأمنية.

باختصار، فاش certificate pinning إجراء أمني مهم، خاصها تستعمل بالاقتران مع ممارسات أمنية أخرى باش نتأكدو بأن API callsاللي كتستعمل Axios ولا أي مكتبة أخرى تكون آمنة قدر الإمكان.

Figure 6: Round 2 of an Illustrative example from Darija Deita dataset.

Human اوكي، كيفاش نقادو axios باش نستعملو certificate pinning ؟

Assistant

باش تصاوب Node.js ف `https` module ، ممكن تستعمل الcertificate pinning باش تخدم Axios باش تكونفيگوري معرفين مسبقًا. ها هو pins واحد المجموعة ديال against ديال السيرڤر certificate لي كيڤاليدي ال HTTPS agent واحد ال واحد المثال كيفاش دير هادشي:

```javascript
const https = require('https');
const axios = require('axios');

// Define the list of pins as an array of SHA-256 hashes
const pins = ['sha256/xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx',
'sha256/yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy'];

// Create a custom HTTPS agent with pin validation
const agent = new https.Agent({
  pinning: true,
  checkServerIdentity: function(host, cert) {
    const certPins = cert.getPublicKey().getFingerprints();
    for (let pin of pins) {
      if (certPins.has(pin)) {
        return undefined; // Certificate is valid
      }
    }
    return new Error('Certificate pin validation failed');
  }
});

// Use the custom HTTPS agent with Axios
const instance = axios.create({
  httpsAgent: agent
});

// Make a secure API call with Axios
instance.get('https://example.com/api/data')
  .then(response => {
    console.log(response.data);
  })
  .catch(error => {
    console.error(error);
  });
```

فهاد المثال، أول حاجة كنديفينيو واحد ال array ديال SHA-256 hashes لي كيمثل اللائحة ديال ال pins لي بغينا نڤاليديو ال certificateديال السيرڤر against ها. من بعد كنصاوبو واحد ال HTTPS agent باستعمال الconstructor `https.Agent` ، وكنحطو ال option `pinning` ل `true` باش نفعل ال.pin validation

من بعد، كنديفينيو واحد ال function `checkServerIdentity` لي كتعيط ليها فاش كيوصل ال certificate ديال السيرڤر. هاد الدالة كتجبد ال SHA-256 fingerprints ديال ال public key ديال ال certificate وكتقارنهم مع اللائحة لي معرفة مسبقًا ديال ال .pins إلا تلاقى شيmatch ، الدالة كترجع`undefined` ، وكاتبين باللي ال certificate صحيح. إلا مالقاتش شيmatch ، الدالة كترجع واحد ال object `Error` بواحد ال message لي كيبين باللي ال pin validation فشل.

أخيرا، كنصاوبو واحد ال instance ديال Axios باستعمال الmethod `axios.create()` ، وكنحطو ال `httpsAgent` option ال HTTPS agent ديالنا. ديك الساعة ممكن نصاوبو واحد ال secure API call باستعمال ال `instance.get()` method، لي غيستعمل ال agent ديالنا باش يڤاليدي ال certificate ديال السيرڤر against ال pins ديالنا لي معرفين مسبقًا.

لاحظ باللي هادا غير واحد المثال كيفاش تكونفيگوري Axios باش تخدمcertificate pinning ، وكاينين طرق أخرى باش تنفذ هاد ال feature على حساب ال specific use case وال environment ديالك.

Figure 7: Round 3 of an Illustrative example from the Darija Deita dataset.

Human سير ل URL اللي جاي باش تبدا التحليل الشامل ديال التطور العالمي ديال التطعيم ضد :COVID-19
https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/vaccinations/vaccinations.csv.
اقترح منهجية مبتكرة ديال تصوير البيانات (data visualization) اللي كتمثل مجموعة البيانات المعقدة بشكل مفهوم.

Assistant
باش تصاوب شي تصويرة فعالة وعندو معنى، خاص تكون عندك فهمة واضحة للمعطيات والقصة اللي كتعاود. ها بعض الطرق المبتكرة لتصور المعطيات اللي تقدر تفكر فيها بالنسبة لهاد ديال التلقيح ضد dataset COVID-19:

1. **Time-Series Analysis**:
هادي تقدر تكون الخطوة الأولى فين كتحلل التقدم ديال التلقيح مع مرور الوقت. هادشي ممكن تمثلو بLine Plots ، فين محور السينات (x-axis) كيمثل الوقت، ومحور العينات (y-axis) كيمثل عدد التلقيحات. خطوط مختلفة تقدر تمثل بلدان مختلفة.
2. **Geo-mapping**:
ممكن تستعمل Geographical Heatmaps باش تبين التغطية ديال التلقيحات ضد COVID-19 فالعالم. المناطق اللي فيها الظلال غامقة كتمثل أعداد كبيرة ديال التلقيحات. هادشي غايعطي نظرة واضحة كيفاش المناطق المختلفة كتمشي فالتقدم ديال التلقيحات.
3. **Comparison Dashboard**:
دخّل bar charts ولا pie charts اللي كتقارن نسب التلقيح بين بلدان مختلفة، وكيبينو الدول اللي قريبة من مناعة القطيع (herd immunity). وزيادة على هادشي، stacked bar ولا line charts ممكن تستعمل باش تبين النسبة ديال الجرعة الأولى إلى الجرعة الثانية ديال التلقيحات لكل بلد.
4. **Trend Analysis**:
بين المعدل باش كل دولة كتلقح السكان ديالها. لهادشي، استعمل خط بياني فين محور السينات (x-axis) كيشير للوقت ومحور العينات (y-axis) كيبين التلقيحات اليومية. خطوط مختلفة تقدر تدل على بلدان مختلفة.
5. **Correlation Heatmap**:
هادا غادي يساعد في فهم كيفاش شي parameter (بحال مجموع التلقيحات ولا الناس اللي تلقحو) مرتبط ب parameters آخرى (بحال مجموع السكان ولا النسبة المئوية ديال السكان).
6. **Aging Analysis**:
عرض الأعمار ديال الناس اللي كياخدو اللقاح بدمج معطيات التلقيح مع معطيات ديموغرافية.
7. **Vaccine Type Analysis**:
في حالة إلى ال dataset كيتضمن النوع ديال اللقاح اللي تعطى، ممكن تصاوب dashboard كيمثل اللقاحات الأكثر استخداما فالعالم، ولا شي robust chart كيبين استخدام اللقاحات فبلدان مختلفة.

تذكر، التصورات خاص ديما تكون تفاعلية باش تعاون المستخدم. أدوات بحال Tableau, Power BI, Python libraries بحال Matplotlib, Seaborn, Plotly, ولا JavaScript libraries بحال D3.js ممكن تستعمل باش تصاوب هاد التصورات القوية.

Figure 8: Illustrative example 1 from the Darija TULU dataset.

Human كتب دالة لي ماشي غي تبدل كاع ليسباس فالسلسلة النصية ب ''%20''، ولكن حتى تحيد أي ليسباس كاين فالأول و الآخر قبل ما تبدل. إضافة على هادشي، الدالة خاصها تخدم مع ليسباسات المتتابعين، يعني تبدلهم ب ''%20'' وحده. الدالة خاصها تكون حساسة لحالة الأحرف(case-sensitive) ، يعني تفرق بين الحروف الكبيرة و الصغيرة.
باش نزيدو الصعوبة، الدالة خاصها تبدل أي تواجد للحروف'a' ، 'b'، 'c'بالقيم ASCII ديالهم بصيغة hexadecimal. التبديل خاصو يكونcase-sensitive ، يعني'A' ، 'B'، 'C'خاصهم يتبدلو بقيم مختلفة على'a' ، 'b'، 'c'. الدالة خاصها تخدم مع التكرارات المتتالية ديال هاد الحروف و تبدلهم بقيمة hexadecimal وحدة مناسبة.
علاوة على ذلك، الدالة خاصها تقدر تخدم مع سلاسل نصية الطول ديالها كيوصل ل 1 مليون حرف. الدالة خاصها تكون مهيئة للسرعة و ما تستعملش شي دوال مدمجة للتلاعب بالسلاسل النصية.

Assistant تنفيذ بحال هاد ال 'functionality' بكفاءة، خصوصا بلا ما نعتمدو على ال built-in string manipulation functions، كيتجاوز القدرات ديال شي لغة interpreted بحال Python ولا JavaScript لي كتستعمل عادة في 'interactive mode'.
ولكن، ها واحد الطريقة ب Python مع شوية ديال ال built-in functions باش نفهمو مزيان:

```python
def replace_chars(string):
    string = string.strip()
    nspaces = 0
    result = ''
    ascii_hex = {'a': '%61', 'b': '%62', 'c': '%63',
    'A': '%41', 'B': '%42', 'C': '%43'}

    for char in string:
        if char.isspace():
            nspaces += 1
            if nspaces == 1:
                result += '%20'
        else:
            nspaces = 0
            if char.lower() in ascii_hex:
                if not result.endswith(ascii_hex[char]):
                    result += ascii_hex[char]
            else:
                result += char

    return result
```

هادشي ماجاوبش على السؤال ديالك نيشان حيت كيستعمل الbuilt-in string manipulation functions ، بسبب التعقيد اللي غايكون ضروري باش نتفاداو كاع هاد الدوال.
لغة بمستوى أدنى بحال C++ أو Rust غادي تكون مناسبة كثر لهاد المستوى ديال ال string manipulation و optimization. بشكل عام، بحال هاد ال) constraint بحال عدم استخدام ال built-in functions والتعامل مع سلاسل نصية فيها حتى لمليون حرف مع تبديلات معينة (ماشي بالضرورة يكون مطلوب في تطوير التطبيقات، و كيشبه لشي مهمة متخصصة و بمستوى منخفض ديال التعامل مع النصوص.
أيضا، مهمة التعامل مع سلسلة نصية فيها حتى لمليون حرف كتطلب تفكير عميق في إدارة الذاكرة(memory management) ، شي حاجة لي اللغات ذات المستوى الأدنى كتكون مؤهلة ليها أحسن. إضافة على ذلك، تقنيات ال optimization بحال string pooling، استخدام هياكل بيانات محددة (eg. ropes or gap buffers) كيتدارو بشكل أحسن بلغات ذات مستوى منخفض.
إذن، هادا موضوع متقدم كيتطلب معرفة بلغة صديقة للcompiler ، فهم الASCII character sets ، و ال memory management.ننصح باستشارة مهندس برمجيات (software engineer) أو مطور متخصص في تحسين الأداء و ال string manipulation.

Figure 9: Illustrative example 2 from the Darija TULU dataset.

Figure 10: Training loss (per step) for GEMMAROC-27B over 1 300 global steps, showing a rapid decline from 3.5 down to 1.3 as the model learns.

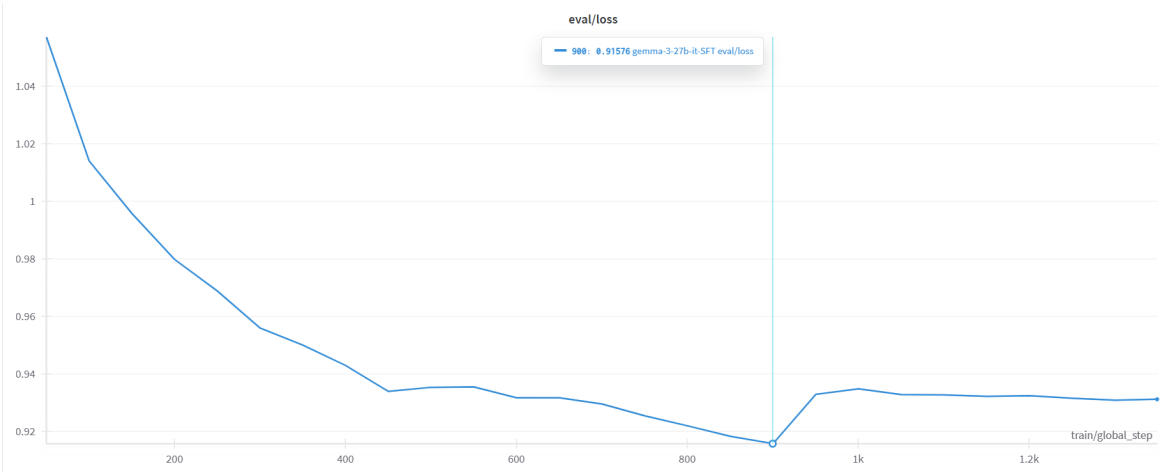

Figure 11: Evaluation loss at each validation checkpoint (every 50 steps), reaching its minimum (0.915) at global step 900 (≈2 epochs), indicating that after just two passes over the data we obtain our most powerful fine-tuned model.