

Enemray: Toward Capable Language Models for Hassaniya

Cheikh Ahmed

Abstract

We introduce Enemray, a Hassaniya-centric language model that enables general-purpose interaction in Hassaniya. Enemray is trained around a stability–plasticity objective: acquire strong Hassaniya linguistic and cultural competence while preserving the general reasoning, multilingual, instruction-following, and safety behaviors of a capable instruction-tuned model. The development pipeline separates language acquisition from behavioral specialization. A separately assembled continual-pretraining corpus provides broad exposure to natural Hassaniya and Mauritanian text; layer-selective continual pretraining learns a compact language-specific parameter update; that update is transferred into the instruction-tuned parameter space; and supervised post-training develops conversational, cultural, literary, task-oriented, and cross-lingual behavior. The supervised corpus integrates selected public Hassaniya and Mauritanian resources with a substantially larger body of newly collected, reconstructed, curated, and constructed instruction data, while policy-generated replay provides a retention signal from the reference model’s own behavior distribution. The resulting collection is substantially larger and broader in purpose than existing Hassaniya text resources. In evaluation, Enemray achieves the strongest English → Hassaniya translation among the compared open and proprietary models and the highest overall score on Mauritanian translation error detection, while retaining most of the general capabilities of its instruction-tuned base model on mathematical reasoning, knowledge, code generation, and function calling. This report describes the motivation, data construction, model design, training methodology, and evaluation of Enemray.

Keywords: Hassaniya Arabic; low-resource language modeling; continual pretraining; language adaptation.

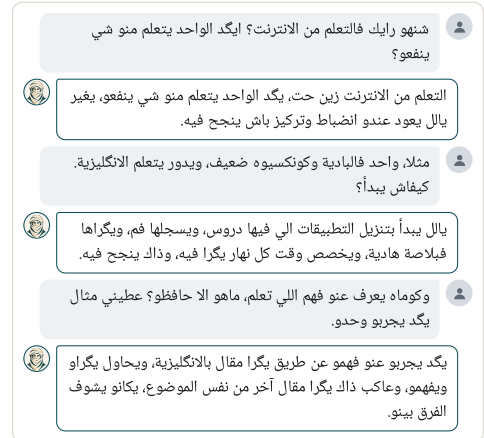
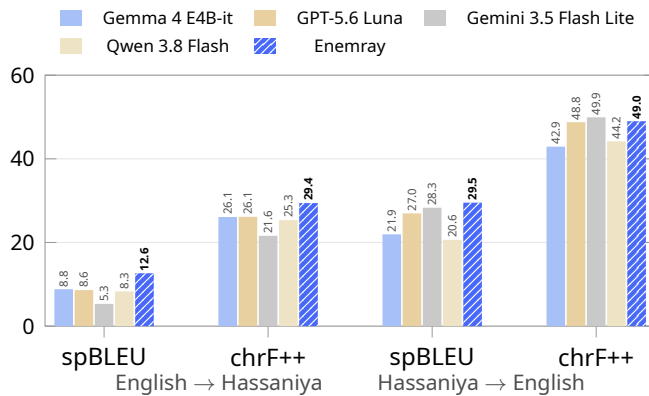


Figure 1. Left: English ↔ Hassaniya translation on the Alexandria (Mauritania) test set, compared with open and proprietary models. Right: a three-turn conversation with Enemray in Hassaniya (unedited model output).

1. Introduction

1.1 Why Hassaniya

Hassaniya (*Hassāniyya*) is a western Arabic variety spoken across a contiguous region of the western Sahara and Sahel, with its largest speaker population in Mauritania, where it functions broadly as a lingua franca. Its historical development reflects sustained contact with Classical and Modern Standard Arabic, Berber languages, and, in modern usage, French and other regional languages [1, 2]. These contact dynamics are visible in Hassaniya phonology, morphology, lexicon, register, and patterns of code-switching, making it a linguistically distinct modeling target [1].

Hassaniya also carries a substantial oral and literary tradition. UNESCO’s documentation of the Moorish epic *T’heydinn* describes a body of dozens of poems in Hassaniya that preserves collective memory, social values, and historical narratives [3]. This cultural role motivates a model spanning everyday conversation and translation alongside local forms of expression, heritage knowledge, and poetic structure.

Despite this linguistic and cultural breadth, Hassaniya remains sparsely represented in computational resources. Earlier large-scale Arabic dialect identification work excluded Mauritania because sufficient dialectal Twitter data could not be collected [4]. More recent resources have expanded coverage through sentiment data, English–Hassaniya parallel corpora, community datasets, culturally grounded Arabic instruction, multi-country dialectal translation, task-oriented dialogue, proverb collections, OCR text, and speech [5–13]. The resulting ecosystem is valuable, but it is distributed across distinct tasks and modalities rather than organized as a broad training distribution for a general-purpose Hassaniya language model.

The central objective of Enemray is to make Hassaniya a first-class interaction language for a general-purpose language model: a language in which the model can reason, explain, translate, converse, write, and engage with culturally specific knowledge and expressive forms.

1.2 The adaptation problem

Specializing a strong multilingual model to an underrepresented language creates a stability–plasticity problem. The model must acquire new linguistic structure from a comparatively concentrated target-language corpus, while retaining capabilities learned from much broader pretraining and post-training distributions. Naive continued training can improve target-language modeling while moving parameters away from behaviors that support reasoning, instruction following, multilinguality, coding, and safety. This challenge is particularly acute for assistants that must combine target-language competence with broad general-purpose reasoning and interaction capabilities.

Enemray addresses this problem by separating *language acquisition* from *behavioral specialization*. First, a language-specific update is learned from a base model through layer-selective continual pretraining (CPT). Second, this update is transferred into the corresponding instruction-tuned model through parameter-space composition. Third, supervised post-training, implemented as supervised fine-tuning (SFT), binds the acquired Hassaniya competence to interaction behaviors. General-capability replay is included within the third stage to maintain exposure to broader assistant behavior during supervised post-training. Enemray implements this decomposition using Gemma 4 E4B. The design follows the broader observation that language addition, parameter-efficient adaptation, and post-training retention are related but distinct optimization problems [14–16].

1.3 Contributions

The technical contributions of this report are fourfold:

- **Large-scale Hassaniya data construction.** We construct two complementary data streams: a separately assembled continual-pretraining corpus for broad language acquisition, and a Hassaniya supervised corpus that integrates selected public resources with a substantially larger body of Enemray-constructed instruction data. Together they cover naturally occurring language, Mauritanian journalism and cultural writing, poetry, translation, grammar, role-play, task-oriented dia-

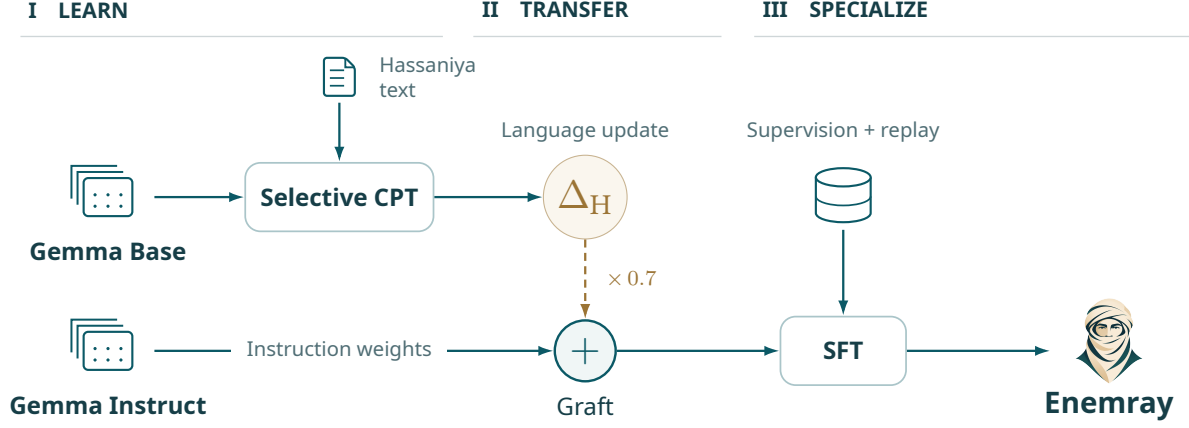


Figure 2. Enemray training pipeline. Selective continual pretraining learns a language update from natural Hassaniya text. The scaled update is added to the instruction-tuned checkpoint, which is then specialized using supervised examples and replay. Solid arrows indicate the main data or model flow; the dashed arrow denotes parameter transfer.

logue, and cultural knowledge. The resulting collection supports a broader training objective than the individual task-specific resources summarized in this report.

- **Layer-selective language acquisition.** We use LayRA-style continual pretraining to place the primary language-learning update in selected early and late layers, reducing the number of trainable parameters while keeping intermediate-layer weights fixed [14].
- **Language capability transfer.** We represent the acquired Hassaniya competence as a parameter update and compose it with the corresponding instruction-tuned model. This follows the task-vector view that fine-tuning-induced parameter differences can encode transferable behavior in weight space [17].
- **Replay-regularized post-training.** We combine Hassaniya-centric supervised learning with offline general-capability replay. The replay component is motivated by recent evidence that training on model-generated, policy-aligned trajectories can reduce forgetting during continual supervised fine-tuning [16].

The full development path is summarized in Figure 2. Each stage answers a different requirement: acquiring the language, transferring that acquisition to an interactive model, shaping Hassaniya behavior, and stabilizing broad capabilities during specialization.

2. Model Design

2.1 Architectural foundation

Enemray adapts the Gemma 4 E4B base and instruction-tuned checkpoints. The backbone is a dense multimodal model with approximately 8 billion parameters including embeddings and approximately 4.5 billion effective parameters. The distinction reflects per-layer embedding tables rather than mixture-of-experts activation [18]. The text decoder contains 42 layers, with hidden dimension 2,560, feed-forward dimension 10,240, and a vocabulary of 262,144 entries. The decoder repeats a block of five sliding-attention layers followed by one full-attention layer seven times, and the configured context capacity is 131,072 tokens [19].

The complete multimodal model is loaded, while adaptation is restricted to attention and feed-forward projections in the text decoder. Vision and audio encoders, multimodal projectors, per-layer embedding components, token embeddings, and the language-model head remain frozen.

Table 1. Inherited Gemma 4 E4B configuration [18, 19].

Property	Specification
Parameter scale	Approximately 8B total; 4.5B effective
Text-decoder layers	42
Hidden / feed-forward dimension	2,560 / 10,240
Vocabulary	262,144 entries
Attention heads / KV heads	8 / 2
Attention pattern	Seven blocks of five sliding and one full-attention layer
Attention-layer totals	35 sliding and 7 full-attention layers
Sliding window	512 tokens
Architectural context	131,072 tokens
Per-layer embedding contribution	256 dimensions
KV-state sharing	Final 18 text layers
Inherited input modalities	Text, image, audio
Adaptation modality	Text only

2.2 Layer-selective adaptation

Following LayRA, which concentrates language adaptation in the early and late layers of a transformer while leaving its intermediate representations unchanged [14], Enemray restricts continual pretraining to the first ten and the final two layers of the text decoder:

$$\mathcal{L} = \{0, 1, \dots, 9\} \cup \{40, 41\}. \quad (1)$$

Continual pretraining therefore updates 12 of the 42 decoder layers, and the remaining 30 intermediate layers retain their pretrained weights. Within each selected layer, low-rank adapters are attached to the query, key, value, and output projections of the attention block and to the up- and down-projections of the feed-forward block, while the feed-forward gate projection is left frozen. All adapters use rank $r = 16$ and scaling parameter $\alpha = 32$ [20].

Continual pretraining thus adapts 72 weight matrices in total. During supervised post-training, adaptation is extended to every layer of the decoder and to all attention and feed-forward projections, including the gate projection, yielding 294 adapted weight matrices.

2.3 Language capability transfer

Let θ_{base} denote the base checkpoint, θ_{post} its instruction-tuned counterpart, and $\theta_{\text{base,H}}$ the base checkpoint after Hassaniya continual pretraining. For each adapted weight matrix W , continual pretraining learns a low-rank update

$$\Delta W_{\text{H}} = \frac{\alpha}{r} B A, \quad (2)$$

where A and B are the learned low-rank factors. Because all other parameters remain unchanged, the language update of the full model is

$$\Delta_{\text{H}} = \theta_{\text{base,H}} - \theta_{\text{base}}, \quad (3)$$

which equals ΔW_{H} for each adapted matrix and zero elsewhere. This update is then added to the instruction-tuned checkpoint:

$$\theta_{\text{H,post}} = \theta_{\text{post}} + \gamma \Delta_{\text{H}}, \quad \gamma = 0.7. \quad (4)$$

We refer to the resulting model $\theta_{\text{H,post}}$ as the grafted checkpoint.

This formulation follows task arithmetic, which interprets differences between fine-tuned and pre-trained weights as reusable directions in parameter space [17], and is closely related to chat-vector

Table 2. Publicly documented Hassaniya or Mauritanian-Arabic resources identified during corpus development. Corpus-wide sizes are shown when a source does not publish a Hassaniya-only count. “SFT” indicates incorporation into Enemray supervised post-training only.

Resource	Type	Publicly reported scale	Enemray use
HASSANIYA [5]	Sentiment	2,000 labeled Facebook comments	–
HASSANIYA-DTCD [22]	Classification	1,851 labeled text records	–
SMART-9K-H-CORP [6]	Parallel text	9,000 English–Hassaniya sentence pairs	–
DAH [7]	Parallel text	3,002 English–Hassaniya rows in Arabic script and Latin transliteration	SFT
PALM [8]	Cultural instruction	1,298 Mauritania instructions, including 294 dialect-category instructions	SFT
Alexandria [9]	Multi-turn translation / dialogue	107K samples overall across 13 Arab countries and 11 domains, including Mauritania	SFT
AI-for-RIM [10]	Translation / dialogue	About 4,430 translation pairs and 594 support turns	SFT dialogue
Amthal Hassaniya [11]	Proverbs	319 Hassaniya proverb examples with explanations	SFT
Hassaniya Stories OCR [12]	OCR / cultural text	552 image–text rows	–
Casablanca [13]	Speech	About 48 hours overall across eight Arabic dialects, including Hassaniya	–
Hassaniya Speech Dataset [23]	Speech	294 audio–text rows	–
Peace Corps Hassaniya material [24]	Language learning	Hassaniya–English instructional text and phrases	–
Basic English–Hassaniya Dictionary [25]	Lexical reference	More than 8,000 words, phrases, and sentences	–
Hassaniya scripture resources [26, 27]	Parallel text	Verse-aligned Hassaniya scripture; edition-dependent	SFT

transfer, in which weight differences between base and instruction-tuned models are combined across checkpoints that share a common origin [21]. The transfer assumes that θ_{post} is derived from θ_{base} and shares its parameterization, so that an update learned relative to the base model remains meaningful in the neighborhood of the instruction-tuned weights. The coefficient γ controls the strength of the transferred update; we use $\gamma = 0.7$ as a fixed setting and leave its systematic selection to future work. Subsequent supervised training specializes the grafted checkpoint for Hassaniya interaction.

3. Training Data

Data is the central constraint in Hassaniya language modeling. Public resources have expanded in recent years, but they remain distributed across narrow tasks such as sentiment classification, translation, cultural instruction, dialogue, speech, proverbs, and language learning. Enemray therefore treats the public ecosystem as a starting point rather than as a complete language-model training distribution. The resulting data pipeline separates two objectives: broad language acquisition from independently assembled natural text during continual pretraining, and behavioral specialization from a larger Hassaniya supervised mixture that incorporates selected public resources together with additional Enemray-constructed and curated data.

3.1 Existing Hassaniya data landscape

Table 2 summarizes publicly documented resources containing Hassaniya or explicitly labeled Mauritanian Arabic that we identified during corpus development. The resources are not directly comparable—they differ in modality, task, annotation, and unit of measurement—but together they show both the growing value and the fragmentation of the available ecosystem. The final column indicates which public resources are incorporated into Enemray; all of them are used in supervised post-training.

The ecosystem is useful but task-specific. SMART-9K-H-CORP provides substantial translation supervision, Casablanca contributes speech coverage, and newer multidialectal efforts such as PALM and Alexandria include Mauritania within broader Arabic collections [6, 8, 9, 13]. None of these resources alone supplies the broad natural-text and general-purpose conversational distribution needed for a Hassaniya generative model. Enemray therefore integrates selected public supervision while constructing a substantially larger training distribution around natural Hassaniya text and general assistant behavior.

3.2 Building the Enemray training corpus

The Enemray training data comprise two complementary streams that serve distinct objectives. The first is a continual-pretraining corpus assembled from naturally occurring and long-form Hassaniya text, including social narratives, reconstructed long-form narratives, Mauritanian news and articles, cultural essays, poetry, and language-learning material. Its purpose is to expose the model to the distribution of the language itself. The second is a supervised corpus designed to shape assistant behavior. It integrates selected public resources from Table 2 with a larger body of Enemray-constructed data spanning cultural knowledge, dialogue, translation, linguistic refinement, practical interaction, and Hassaniya literary forms.

Table 3 summarizes the size of each stream after cleaning. Because the two streams are organized around different training units, the continual-pretraining corpus is measured in documents and the supervised corpus in conversations, and both are additionally reported in tokens of the Gemma 4 tokenizer. The supervised stage also includes a small general-capability replay set. Although it is not Hassaniya data, it is listed here to present the complete supervised mixture, and its role is discussed with the training methodology in Section 4.

Table 3. Size of the Enemray training data after cleaning. Token counts are computed with the Gemma 4 tokenizer over message content, excluding system prompts and conversational formatting.

Component	Size	Gemma 4 tokens
Continual-pretraining corpus	11,629 documents	20.9M
Hassaniya supervised data	46,259 conversations	3.0M
General-capability replay	273 conversations	0.4M
Complete supervised mixture	46,532 conversations	3.4M

The supervised mixture contains 46,532 conversations with 50,273 assistant turns. Since each assistant turn is treated as a separate training example, the number of assistant turns corresponds to the number of supervised examples used during post-training, whereas the number of conversations reflects the size of the underlying dialogue collection.

Within this design, public datasets constitute one component of the supervised stage rather than its foundation: data constructed for Enemray account for roughly two thirds of the supervised response tokens, while public resources contribute the remaining third. In scale, the continual-pretraining corpus alone exceeds the individual public text resources in Table 2, which are typically measured in thousands of sentence pairs or instructions, by more than an order of magnitude.

3.3 Data construction recipes

Beyond integrating public resources, Enemray relies on several explicit construction recipes designed around the forms in which Hassaniya is naturally available. These recipes preserve the distinction between naturally occurring text, machine-assisted derived supervision, and deliberately constructed normalization data. Figure 3 summarizes the main recipe families and how they feed either the continual-pretraining stream or the Hassaniya supervised mixture.

Natural web text for continual pretraining. The continual-pretraining stream is built primarily through web collection from many Mauritanian online sources. We collect naturally occurring Has-

saniya posts, social narratives, stories, articles, news, and related text, then organize the retained material as document-level causal-language-modeling examples. This stream is intentionally kept separate from the supervised public datasets in Table 2: its role is exposure to the language distribution itself rather than instruction following. Social posts and narratives form the largest component because they provide substantial naturally occurring colloquial Hassaniya, while journalism, long-form articles, cultural writing, and literary material broaden register coverage.

Context-guided pseudo-parallel translation. Web posts and stories collected for continual pre-training are also used to build supervised translation data. Each document is split into short and medium Hassaniya segments, which become the source side of the translation pairs. A short segment taken on its own is often ambiguous: it may refer to people or events mentioned earlier, omit information that is clear from the surrounding text, or use a local expression whose meaning depends on the situation. For this reason, Gemini [28] does not translate segments in isolation. It receives the full source document together with the selected segment, and produces an English rendering of that segment only, using the document to resolve the narrative situation, referents, and dialectal expressions. Because the English side is produced by a model, the resulting pairs are treated as *machine-generated pseudo-translations* rather than human-authored gold translations. Before inclusion, each rendering is checked by a human against the Hassaniya segment and its context.

Context-guided conversation construction. User–assistant conversations are built from the same natural Hassaniya documents. Generating dialogue from an isolated sentence tends to produce generic exchanges that are only loosely related to the source. Instead, Gemini receives the source document and writes conversational turns grounded in its content, so that the exchange reflects the situations, people, and topics that appear in the original text. As with the translation data, each candidate conversation is verified by a human before it enters the supervised mixture. A single source document can therefore contribute both pseudo-parallel translation pairs and contextual assistant conversations, while both types of supervision remain tied to naturally occurring Hassaniya.

Correction and Hassaniya normalization. Correction data is constructed in the reverse direction: a clean Hassaniya sentence is kept as the assistant target, while a controlled noisy version is presented as the user input. The corruption pipeline includes rule-based character, spacing, punctuation, spelling, and local word-order noise; machine-translation (MT) and back-translation noise that can introduce translationese or less-natural phrasing; and conservative substitution of selected Hassaniya forms with Modern Standard Arabic (MSA) or Moroccan Darija alternatives. The goal is broader than spelling repair: the model is trained to normalize noisy, mixed, or model-like text back into natural Hassaniya while preserving meaning. Generated correction examples are built only from the training split of the clean translation targets; held-out translation examples and Alexandria translation evaluation data are excluded from this augmentation.

Poetry and literary curation. Because poetry and oral expression are central to Hassaniya cultural and linguistic coverage, we also select substantial Hassaniya poetic and literary material from additional sources rather than relying only on generic Arabic corpora. Raw literary text contributes language exposure where appropriate, while structured examples are used to teach supervised literary and cultural behavior, including Hassaniya poetic forms such as Aagel, Lghna, and Tebraa. This provides lexical, stylistic, figurative, and culturally grounded patterns that are comparatively sparse in ordinary web prose.

3.4 Continual-pretraining corpus

The continual-pretraining corpus exposes the base model to natural text before assistant specialization. The clean corpus contains 11,629 documents totaling approximately 20.9 million Gemma 4 tokens (9.5 million words). Documents are typically long-form, with a median length of 581 tokens. Its source domains are summarized in Table 4.

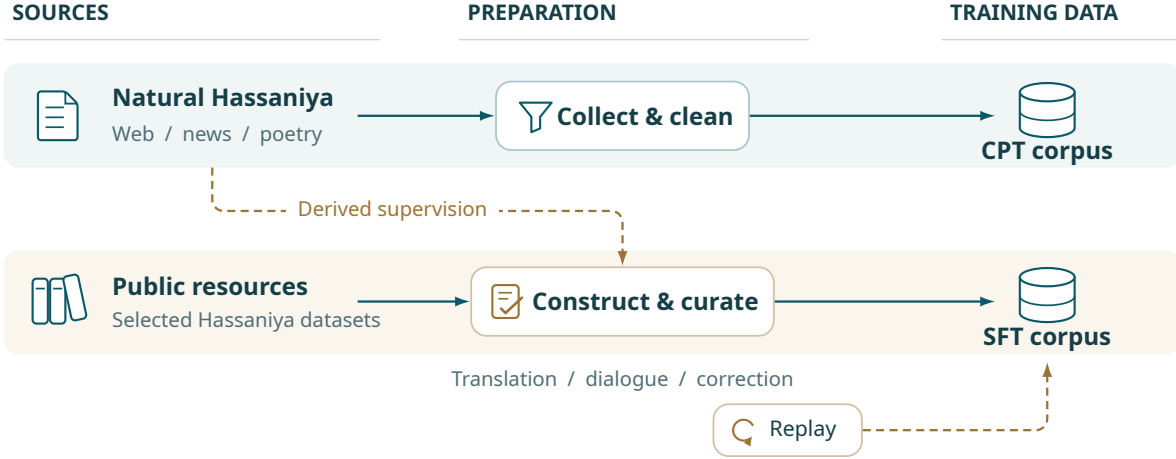


Figure 3. Two complementary training-data streams. Natural text supplies CPT and material for derived supervision. Public resources join constructed examples through curation; replay enters the final SFT mixture separately. Dashed arrows identify derived supervision and replay inputs. Construction includes context-guided generation, human verification, normalization, and literary curation, as described in the text.

Table 4. Source domains of the continual-pretraining corpus.

Source domain	Contribution
Standalone social-media narratives	Colloquial writing and everyday discourse
Reconstructed long-form narratives	Extended narrative and contextual continuity
Mauritanian news media	Contemporary events and expository language
Mauritanian long-form articles	Longer factual and topical discussion
Cultural essays	Local knowledge and cultural expression
Poetry and language-learning texts	Literary forms and linguistic explanation

The mixture is intentionally register-diverse. Social narratives provide everyday dialectal discourse, dialogue, interpersonal relations, narrative progression, colloquial lexical choice, and naturally varying orthography. Journalism and long-form articles extend coverage toward factual, expository, and more formal registers. Cultural essays and language-learning material add local knowledge and pedagogical structure, while poetry introduces expressive forms that are culturally salient in Hassaniya and difficult to recover from general Arabic corpora alone [3].

3.5 Hassaniya supervised corpus

The Hassaniya supervised data comprise 46,259 conversations and 49,964 assistant turns, with approximately 3.0 million tokens of user and assistant content, of which 1.5 million tokens form the assistant responses used as training targets. Selected public datasets contribute high-value forms of supervision that are difficult to recover reliably from raw text alone. DAH contributes English–Hassaniya parallel examples; the Mauritanian portion of Alexandria adds multi-domain conversational translation; PALM contributes country-specific cultural and dialectal instructions; selected AI-for-RIM customer-support dialogues add practical multi-turn interaction; Amthal Hassaniya contributes proverb explanation and idiomatic grounding; and Hassaniya scripture parallels add further cross-lingual lexical and syntactic coverage [7–11, 26, 27].

Beyond these public sources, Enemray data extend supervision across five capability families: **cultural and contextual grounding**, including Mauritanian heritage, history, local knowledge, and question answering; **poetry and oral tradition**, including Aagel, Lghna, and Tebraa; **situational conversation**, including role-play and everyday social scenarios; **cross-lingual communication**, including English–Hassaniya dialogue and translation; and **linguistic refinement**, including grammar, spelling, phrasing, lexical choice, correction, and stylistic naturalness.

3.6 Curation, separation, and supervision

The data pipeline is organized around the objective of each training stage. Natural text is used for continual pretraining, where it is presented as documents under a causal language-modeling objective, whereas instruction, translation, proverb, dialogue, and parallel-text resources are reserved for supervised post-training. This separation distinguishes learning the distribution of the language from learning how to respond within an interaction.

In the supervised stage, conversations are represented in the native conversational format of the model, and the training signal is restricted to assistant responses. Each assistant turn forms its own training example, with the preceding dialogue serving as context, so that the model learns to produce responses rather than to reproduce prompts. The general-capability replay set is kept separate from the Hassaniya data and is added only to the final supervised mixture, where it serves as a retention signal (Section 4).

Cleaning is deliberately conservative. For the continual-pretraining corpus, it removes invisible formatting characters while preserving every document, including dialectal spelling and informal punctuation. For the supervised corpus, it removes exact duplicate conversations and residual markup from generated responses. These steps ensure structural consistency across the data.

Because supervision is built from the same body of Hassaniya literature, some poems appear both as documents in the continual-pretraining corpus and as responses in the supervised poetry data. The model therefore encounters these texts first as natural language and subsequently as targets of literary generation.

4. Training Methodology

Enemray is trained in three stages that separate language acquisition, language transfer, and supervised post-training (Figure 2). The configuration of each training stage is summarized in Table 5.

4.1 Stage I: Continual pretraining

The base checkpoint is adapted to Hassaniya under the causal language-modeling objective, with low-rank adapters restricted to the layers and projections described in Section 2. Documents from the continual-pretraining corpus are separated by an end-of-sequence token and packed into sequences of 2,048 tokens. Training runs for two epochs, during which the model processes approximately 42 million tokens. Optimization uses 8-bit AdamW with a learning rate of 5×10^{-5} , linear decay after a short warmup, and an effective batch size of 16 sequences.

4.2 Stage II: Language transfer

The trained language adapter is transferred to the instruction-tuned checkpoint by scaling its update with $\gamma = 0.7$ and merging it into the model weights, as defined in Equation 4. The merge is performed in full precision. On a reference input, the merged model reproduces the logits of the unmerged adapted model to within a maximum absolute difference of 1.7×10^{-4} , confirming that merging preserves the behavior of the adapted model.

4.3 Stage III: Supervised post-training

The grafted checkpoint is then trained on the supervised mixture described in Section 3.5, with adaptation extended to all attention and feed-forward projections across the full depth of the decoder (Section 2). Each assistant turn forms a separate training example, in which the assistant response is the target and the preceding dialogue serves as context. Let $T(x)$ denote the target token positions of an example x . The training objective is

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{\sum_x |T(x)|} \sum_x \sum_{t \in T(x)} \log p_\theta(x_t \mid x_{<t}), \quad (5)$$

so that prompt and padding tokens do not contribute to the loss.

Table 5. Training configuration of the continual-pretraining and supervised post-training stages.

Setting	Continual pretraining	Supervised post-training
Maximum sequence length	2,048	8,192
Effective batch size	16	16
Gradient accumulation	1	1
Epochs	2	2
Learning rate	5×10^{-5}	5×10^{-5}
Learning-rate schedule	Linear	Linear
Warmup fraction	0.03	0.03
Weight decay	0.01	0.01
Gradient clipping	1.0	1.0
Optimizer	8-bit AdamW	AdamW
LoRA rank / scaling parameter	16 / 32	16 / 32
LoRA dropout	0	0
Adapted weight matrices	72 in 12 layers	294 in 42 layers
Loss	Causal language modeling	Assistant responses
Random seed	42	42

Conversations are rendered in the native Gemma 4 conversational format with the reasoning mode disabled, and sequences are limited to 8,192 tokens. When a dialogue exceeds this length, the earliest context is removed so that the target response is always retained in full. Training runs for two epochs with AdamW at a learning rate of 5×10^{-5} and an effective batch size of 16, with examples sampled in random order.

The supervised mixture also contains the general-capability replay set, which covers reasoning, coding, multilingual assistance, and instruction following. Replay accounts for approximately 17% of the supervised target tokens and maintains exposure to general assistant behavior while the model specializes in Hassaniya, following evidence that replay of model-generated responses reduces forgetting during continual fine-tuning [16].

4.4 Training setup

All stages are trained on a single NVIDIA H100 80 GB GPU. Model computation uses BF16 precision, while the adapter merge in Stage II is performed in FP32 before the grafted model is stored in BF16. The complete multimodal checkpoint is loaded throughout, and the native Gemma 4 processor and conversational format are retained [29].

5. Evaluation

We evaluate Enemray along three axes: translation between English and Hassaniya, detection and classification of errors in Hassaniya machine translation, and retention of general capabilities after adaptation. All metrics are computed automatically against fixed references or executable tests, without human or model-based judges. Table 6 summarizes the experimental setup.

Table 7. Translation results on the Alexandria (Mauritania) test set, 1,000 examples per direction. Amalaz denotes the direction-specific MarianMT models amalaz-en2ha-micro and amalaz-ha2en-micro. The best score in each row is shown in bold.

	Metric	Amalaz micro	Gemma 4 E4B-it	GPT-5.6 Luna	Gemini 3.5 Flash Lite	Qwen 3.8 Flash	Enemray
English → Hassaniya	spBLEU	5.74	8.83	8.61	5.32	8.30	12.64
	chrF++	21.96	26.09	26.14	21.59	25.34	29.42
Hassaniya → English	spBLEU	13.22	21.94	26.95	28.31	20.62	29.52
	chrF++	32.76	42.91	48.76	49.93	44.18	48.99

Table 6. Evaluation setup. All samples are drawn once with a fixed random seed and shared across models.

Experiment	Data	Samples	Metric
Translation	Alexandria (Mauritania), English → Hassaniya and Hassaniya → English	2,000	spBLEU, chrF++
Error detection	AlexandriaX Subtask 3 (Mauritania), LQM-annotated error spans	274	Exact match F1, overlap F1, class F1
Math reasoning	GSM8K [30]	500	Exact match
Knowledge	MMLU-Pro [31]	500	Accuracy
Code generation	HumanEval [32]	164	pass@1
Function calling	BFCL [33]	500	Call accuracy

Models and decoding. Enemray is compared with its underlying instruction-tuned model, Gemma 4 E4B-it, so that differences reflect the adaptation pipeline. For translation and error detection, we additionally include three proprietary and open models accessed through OpenRouter: GPT-5.6 Luna, Gemini 3.5 Flash Lite, and Qwen 3.8 Flash. For translation, we also report Amalaz, a pair of small direction-specific MarianMT models for English → Hassaniya and Hassaniya → English, which translate the source sentence directly with their default beam-search decoding. Gemma 4 E4B-it and Enemray are decoded greedily with the reasoning mode disabled. The external models are queried with a fixed seed, a low reasoning effort, and the provider’s default sampling configuration. Within each experiment, every model receives the same prompt.

5.1 Translation

Translation is evaluated on the Mauritanian portion of Alexandria [9], a human-translated and reviewed collection of multi-domain dialogues. From its test split, we sample 1,000 parallel dialogue turns covering all 11 domains and evaluate each turn in both directions, yielding 1,000 English → Hassaniya and 1,000 Hassaniya → English examples. Each model receives a neutral translation instruction specifying the source and target languages and asking for the translation only. Outputs are scored against the reference translation with spBLEU, computed with the FLORES-200 SentencePiece tokenizer [34], and chrF++ [35], both implemented in sacreBLEU [36].

Results. Table 7 and Figure 1 report the translation results. Enemray achieves the highest English → Hassaniya scores on both metrics, improving over Gemma 4 E4B-it by 3.8 spBLEU and 3.3 chrF++. In the Hassaniya → English direction, Enemray obtains the highest spBLEU, while Gemini 3.5 Flash Lite achieves the highest chrF++.

5.2 Error detection

The error-detection experiment measures whether a model can recognize what is wrong in a Hassaniya translation, rather than produce or repair one. It follows Subtask 3 of the AlexandriaX-2026 shared task

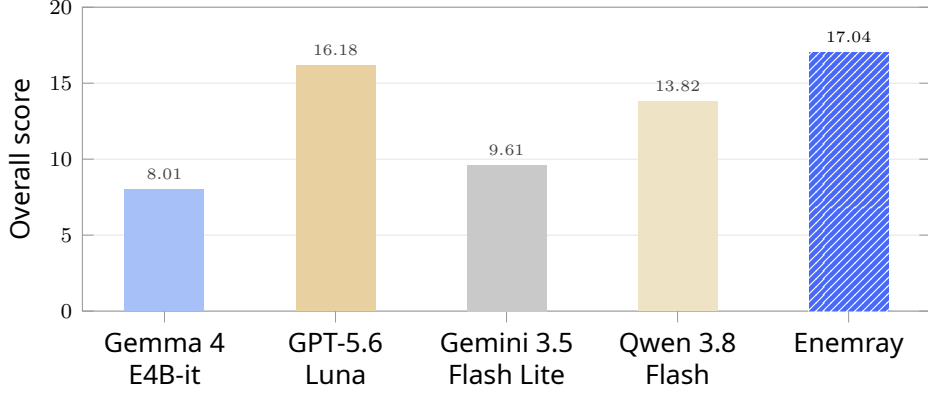


Figure 4. Overall error-detection score on the Mauritanian portion of AlexandriaX Subtask 3, the average of overlap F1 and class F1 (multiplied by 100).

Table 8. Error-detection results on the Mauritanian portion of AlexandriaX Subtask 3 (274 examples, 390 gold error spans). Scores are multiplied by 100, and the best score in each metric row is shown in bold. The lower rows describe prediction behavior; in the gold annotation, erroneous spans cover 19.5% of each translation on average.

	Gemma 4 E4B-it	GPT-5.6 Luna	Gemini 3.5 Flash Lite	Qwen 3.8 Flash	Enemray
Overall	8.01	16.18	9.61	13.82	17.04
Exact match F1	3.39	9.97	6.37	10.03	2.79
Overlap F1	12.21	22.13	13.55	20.39	27.58
Class F1	3.81	10.23	5.66	7.24	6.51
Examples reported as error-free	227	60	174	79	19
Translation marked as erroneous (%)	6.9	23.1	8.2	16.2	46.9

[37], in which a system receives an English sentence and its machine translation into an Arabic dialect, identifies the erroneous spans in the translation, and assigns each span an error category. The underlying data were built with the LQM framework [38]: English sentences derived from conversational dialect transcripts were translated by six Arabic-aware large language models, and native-speaker expert annotators marked the erroneous spans and labeled them with one of six linguistically grounded categories: sociolinguistics, pragmatics, semantics, morphosyntax, orthography and writing conventions, and graphetics. We use all 274 English $\rightarrow$ Mauritanian Hassaniya examples with released annotations, comprising the 244 training and 30 development examples of the shared task, which together contain 390 annotated error spans.

Each model receives the English source, the translation, and the list of the six categories with their LQM definitions, and returns the erroneous spans with their categories in JSON format. Character offsets are obtained by locating each returned span in the translation. Predictions are scored with the official shared-task scorer, which reports exact match F1, where a predicted span must match the gold character offsets exactly; overlap F1, computed over the characters marked as erroneous; and class F1, where predicted and gold spans are paired by their overlap and a pair counts as correct when the categories agree. Following the shared task, the overall score is the average of overlap F1 and class F1. We report results on all 274 examples.

Results. Table 8 and Figure 4 report the error-detection results. The task remains difficult for all models. Adaptation markedly changes the behavior of the base model: Gemma 4 E4B-it reports no error for 227 of the 274 translations, all of which contain at least one annotated error, whereas Enemray reports errors for all but 19 and obtains the highest overall and overlap scores. Its spans are, however, con-

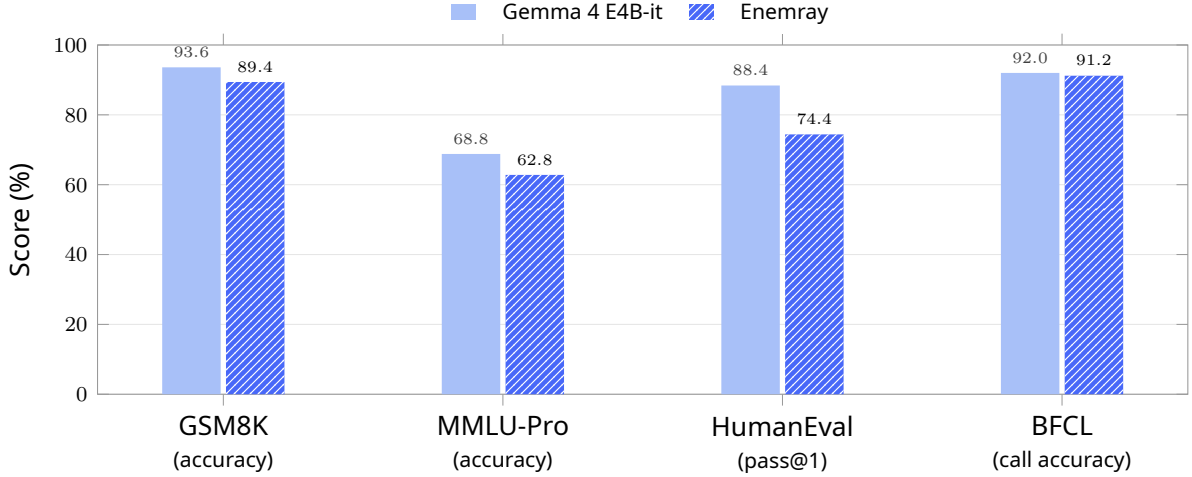


Figure 5. Capability retention of Enemray compared with Gemma 4 E4B-it on GSM8K, MMLU-Pro, HumanEval, and BFCL.

siderably broader than the gold annotations, covering 46.9% of each translation on average compared with 19.5%, which yields the lowest exact match F1. Category assignment is the main weakness across models: although sociolinguistic errors account for 81% of the gold spans, the models predominantly label errors as semantic or morphosyntactic.

5.3 Capability retention

Capability retention compares Enemray with Gemma 4 E4B-it on four established benchmarks, using the same prompts, no system prompt, and greedy decoding for both models.

Mathematical reasoning. We sample 500 problems from the GSM8K test set [30]. The model reasons step by step and states its final answer in a fixed format, and a response is correct when the extracted number equals the reference answer.

Knowledge and reasoning. We sample 500 questions from the MMLU-Pro test set [31], stratified by subject category. Each question has up to ten answer options; the model reasons step by step and concludes with the selected letter, which is extracted following the official MMLU-Pro evaluation. We report accuracy.

Code generation. All 164 HumanEval problems [32] are used. The model completes each function from its signature and docstring, and the completed program is executed against the official unit tests. We report pass@1.

Function calling. We sample 500 single-turn problems from the Berkeley Function Calling Leaderboard [33]: 200 simple, 100 multiple-function, 100 parallel, and 100 irrelevance cases. Tool definitions are provided through the native Gemma 4 function-calling format, and the generated calls are verified with abstract-syntax-tree matching following the official BFCL criteria, which check the function name, required parameters, and parameter values. In the irrelevance category, a response is correct when the model makes no function call.

Results. Figure 5 reports the capability retention results. Enemray retains most of the general capabilities of Gemma 4 E4B-it, with some reduction after adaptation. Function calling is largely unchanged, with BFCL accuracy of 91.2% compared with 92.0%. On GSM8K and MMLU-Pro, Enemray reaches 89.4% and 62.8%, retaining 95.5% and 91.3% of the base model’s accuracy. The largest reduction appears in code generation, where HumanEval pass@1 decreases from 88.4% to 74.4%. Enemray’s solutions tend

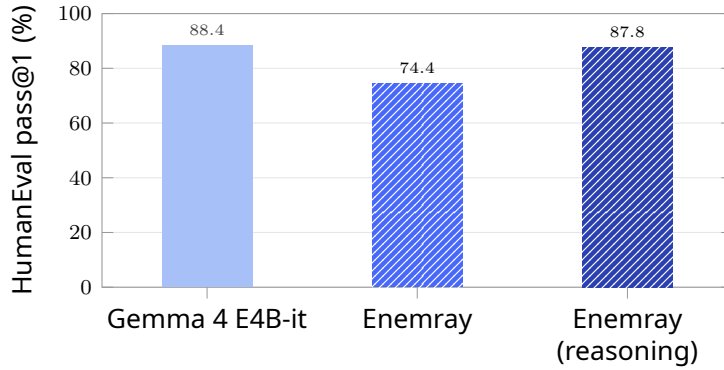


Figure 6. HumanEval pass@1 of Enemray with and without its reasoning mode, compared with Gemma 4 E4B-it. Without reasoning, both models are decoded greedily; with reasoning, Enemray is sampled at temperature 0.8 with up to 8,192 new tokens.

to be shorter and to implement the function directly, whereas the base model more often works through the problem in code comments before implementing it. This suggests that much of the gap reflects how the model approaches the problem rather than a loss of coding knowledge. Consistent with this, when Enemray is evaluated with its reasoning mode enabled (sampling temperature 0.8, up to 8,192 new tokens), its HumanEval pass@1 rises to 87.8%, nearly matching the 88.4% of the base model without reasoning (Figure 6). Although replay accounts for approximately 17% of the supervised target tokens, it consists of only 273 conversations, and coding tasks form a small subset of them. Increasing the number and diversity of policy-generated replay conversations, with broader coverage of coding and reasoning tasks, is a direct way to further improve retention in future versions of Enemray.

6. Limitations and Release

6.1 Limitations

Enemray is trained on a collection that emphasizes natural narratives, social writing, Mauritanian journalism, cultural material, and targeted assistant tasks. Although this collection spans a broad range of registers, coverage of specialized domains such as scientific, legal, and professional writing remains limited, and competence in these domains is correspondingly weaker. Hassaniya is also primarily an oral variety with considerable variation in spelling, code-switching, and proximity to Standard Arabic [1]. Model outputs may therefore alternate between Hassaniya and Standard Arabic forms, and judgments of dialectal naturalness, cultural appropriateness, and poetic form ultimately require native-speaker expertise.

The present adaptation is restricted to text. Although the underlying Gemma 4 E4B checkpoint supports image and audio input, the vision and audio components are not adapted, and the multimodal behavior of Enemray is outside the scope of this report.

6.2 Release of the current checkpoint

The Enemray checkpoint described in this report will not be publicly released. The base Gemma 4 E4B model is distributed under the Apache 2.0 license [18] and does not itself restrict the release of derived models. The constraint arises instead from the supervised training data. Several third-party resources incorporated into the supervised mixture are distributed under terms that limit their use and redistribution. Alexandria is released under a Creative Commons Attribution–NonCommercial license, with access conditioned on an agreement to non-commercial use [9], and PALM is likewise distributed under non-commercial terms [8]. Other resources in the mixture do not state explicit licensing terms, and their use in a publicly distributed model would require permission from the respective rights holders. Because model weights trained on these resources cannot be released without such permissions, the current checkpoint is not made publicly available.

6.3 Toward a releasable Enemray

We are actively working toward a version of Enemray that can be released openly. This effort follows two directions. First, resources whose licensing terms require permission for redistribution or derivative use are being removed from the training mixture. Second, we are developing and curating locally sourced Hassaniya data to replace these resources and to further extend coverage of Hassaniya language, culture, and interaction. The resulting corpus will be used to train a more capable Enemray model, which we intend to release publicly.

Beyond changes to the data, the next version will extend the training pipeline beyond supervised fine-tuning. We plan to incorporate reinforcement learning from human feedback, with preferences provided by native Hassaniya speakers, to better align the model’s responses with the expectations of Hassaniya users in terms of naturalness, cultural appropriateness, and helpfulness. We also plan to simulate native Hassaniya-speaking users, so that the model can be trained and assessed on realistic multi-turn interactions that reflect how Hassaniya speakers actually phrase requests and conduct conversations. Finally, the next version is intended to extend Enemray beyond text by adapting both its audio and image capabilities, allowing the model to understand spoken Hassaniya and to engage with visual content in Hassaniya.

7. Additional Examples

Table 9 presents additional examples of Enemray outputs covering everyday advice, English → Hassaniya translation, and poetry generation.

Table 9. Additional examples of Enemray outputs. Outputs are unedited and were generated with greedy decoding and the reasoning mode disabled.

Task	Input	Enemray output
Everyday advice System: general	السلام عليكم. ندور نعود نخصص شي من وقتي للكراية، غير النهار كامل واحل والليل نعود فتران. شنهو اللي تنصحنى بيه؟	وعليكم السلام. ناصحك تبدأ ببرنامج بسيط. حدد وقت ثابت كل نهار، كيف عاغب صلاة العصر، وخصص فيه ساعة وحدة للكراية.
Translation English → Hassaniya	I waited for you near the market, but my phone ran out of battery. If you arrive before me tomorrow, wait at the same place.	حانيتك عند المرسى، يغير تلفوني اوفات بطاريته. لين اتعود جيت سابغني الصبح، حانيني ف نفس لبلاصة.
Translation English → Hassaniya	Learning a little every day is often more useful than studying for many hours once a month. The important thing is to keep practicing, even when progress feels slow.	تعلم شوي كل نهار غالبا يعود أكثر فائدة من الدراسة لساعات ياسرة مرة كل شهر. المهم هو الاستمرار فالتمرين، حتى لين يعود التقدم بطيء.
Poetry generation System: poetry	اعطيني كاف جديد من أربع تيفلواتن عن واحد بعيد عن أهلو ويتفكد كعدتهم فالليل. يكون فيه الشوگ، وعطيني الكاف وحدو بلا شرح.	الليل يالبال يطول == و آن فالغربة وحدي و اليل أطول و أطول == ما گت نتفكد أهلي

8. Conclusion

This report presented Enemray, a Hassaniya-centric language model built on Gemma 4 E4B. Enemray approaches the adaptation of a capable instruction-tuned model to an underrepresented language as a stability–plasticity problem and addresses it by separating language acquisition from behavioral specialization. Layer-selective continual pretraining learns a compact Hassaniya update from natural text, this update is transferred into the instruction-tuned model through parameter-space composition, and supervised post-training with general-capability replay shapes Hassaniya interaction while maintaining broader assistant behavior. Supporting this pipeline, we assembled a continual-pretraining corpus of natural and long-form Hassaniya text and a supervised corpus that combines selected public resources with a larger body of Enemray-constructed data covering cultural knowledge, poetry, conversation, translation, and linguistic refinement.

The evaluation indicates that this design yields clear gains in Hassaniya while largely preserving

general capabilities. Enemray achieves the strongest English $\rightarrow$ Hassaniya translation among the evaluated models, including the proprietary systems, and the highest spBLEU in the Hassaniya $\rightarrow$ English direction. It also obtains the highest overall score on Mauritanian error detection, although precise span localization and error categorization remain difficult for all models. On general benchmarks, Enemray retains most of the capabilities of its instruction-tuned base model: function calling is largely unchanged, mathematical reasoning and knowledge show modest reductions, and the reduction in code generation is almost entirely recovered when the reasoning mode is enabled.

Several directions remain open. The next version of Enemray will replace resources with restrictive licensing terms by locally sourced Hassaniya data so that the model can be released openly. It will also increase the amount and coverage of policy-generated replay, incorporate reinforcement learning from human feedback with native Hassaniya speakers and simulated Hassaniya-speaking users, and extend adaptation to audio and images. We hope that Enemray and the methodology documented in this report contribute to making Hassaniya a first-class language for modern language technology.

References

- [1] Catherine Taine-Cheikh. Ḥassāniyya arabic. In Christopher Lucas and Stefano Manfredi, editors, *Arabic and Contact-Induced Change*, pages 245–263. Language Science Press, Berlin, 2020. doi: 10.5281/zenodo.3744523. URL <https://langsci-press.org/catalog/book/235>.
- [2] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 5.3: Hassaniyya. Glottolog, 2025. URL <https://glottolog.org/resource/languoid/id/hass1238>.
- [3] UNESCO. Moorish epic t’heydinn. Intangible Cultural Heritage, n.d. URL <https://ich.unesco.org/en/USL/moorish-epic-t-heydinn-00524>.
- [4] Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*, 2020. URL <https://arxiv.org/abs/2005.06557>.
- [5] Med El Moustapha El Arby. Hassaniya dataset. Mendeley Data, 2025. URL <https://data.mendeley.com/datasets/m2swkr2bhx/1>.
- [6] Rahma Boughale, Aichetou Nemin, Mamadou Abdoulaye Aw, Mohamedou Abewa, and Cheikh Seyidi Dedaha Mohamed Jedou. A low-resource english–hassaniya parallel corpus with neural machine translation benchmarks. In *Artificial Intelligence and its Practical Applications*, volume 1916 of *Lecture Notes in Networks and Systems*, pages 187–200. Springer, 2026. doi: 10.1007/978-3-032-22032-5_12. URL https://link.springer.com/chapter/10.1007/978-3-032-22032-5_12.
- [7] Ahlam Abdelkader, Emani Babe, and Oumoukelthoum Sidenna. Dah: Dataset hassaniya. Hugging Face dataset, 2026. URL <https://huggingface.co/datasets/hassan-IA/dah>.
- [8] Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, Abdelrahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, et al. Palm: A culturally inclusive and linguistically diverse dataset for arabic llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.1579/>.
- [9] Abdellah El Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, et al. Alexandria: A multi-domain dialectal arabic machine translation dataset for culturally inclusive and linguistically diverse llms. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, 2026. URL <https://arxiv.org/abs/2601.13099>. arXiv:2601.13099.

- [10] Emin009. Ai for rim: Hassaniya dialect datasets. Hugging Face dataset repository, 2026. URL <https://huggingface.co/datasets/Emin009/AI-for-RIM>. Repository contains separate Arabic–Hassaniya translation and Hassaniya customer-support resources.
- [11] ahmed02mk. Amthal hassaniya: Hassaniya proverb dataset. Hugging Face dataset, 2026. URL <https://huggingface.co/datasets/ahmed02mk/amthal-hassaniya>. 319 Hassaniya proverbs with explanations and usage-oriented supervision.
- [12] Hassan-IA. Hassaniya stories ocr. Hugging Face dataset, 2026. URL <https://huggingface.co/datasets/hassan-IA/hassaniya-stories-ocr>.
- [13] Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou Cheikh Tourad, Rahaf Alhamouri, Rwa Assi, et al. Casablanca: Data and models for multidialectal arabic speech recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21745–21758. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.emnlp-main.1211/>.
- [14] Abraham Toluase Owodunni and Sachin Kumar. Continually adding new languages to multilingual language models. *arXiv preprint arXiv:2509.11414*, 2025. URL <https://arxiv.org/abs/2509.11414>.
- [15] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL <https://arxiv.org/abs/2405.09673>.
- [16] Yan Chen, Taojie Zhu, Meng Zhang, Xin Chen, Jiaqi Huang, Dongyang Xu, and Yizhi Wang. On-policy replay for continual supervised fine-tuning. *arXiv preprint arXiv:2605.29495*, 2026. URL <https://arxiv.org/abs/2605.29495>.
- [17] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. URL <https://arxiv.org/abs/2212.04089>.
- [18] Google DeepMind. Gemma 4 E4B. Hugging Face model card, 2026. URL <https://huggingface.co/google/gemma-4-E4B>. Accessed 12 September 2026.
- [19] Google DeepMind. Gemma 4 E4B: Pinned model configuration. Hugging Face model configuration, 2026. URL <https://huggingface.co/google/gemma-4-E4B/blob/411aa17b749aa952df1359d2dcea73917a544d9a/config.json>.
- [20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL <https://arxiv.org/abs/2106.09685>.
- [21] Shih-Cheng Huang, Pin-Zu Li, Yu-Chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tzong-Han Tsai, and Hung-yi Lee. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. URL <https://arxiv.org/abs/2310.04799>.
- [22] Med El Moustapha El Arby. Hassaniya-dtcd: A new dataset for benchmarking text classification tasks on hassaniya dialect. Mendeley Data, 2025. URL <https://data.mendeley.com/datasets/r5k9ktwr4g/1>.
- [23] Mamadou-Aw. Hassaniya speech dataset. Hugging Face dataset, 2026. URL <https://huggingface.co/datasets/Mamadou-Aw/Hassaniya-speech-dataset>.

- [24] Peace Corps and alvations. Peace corps language resources: Hassaniya. Hugging Face dataset aggregation of Peace Corps language-learning materials, 2026. URL <https://huggingface.co/datasets/alvations/peace-corps>.
- [25] Ahmed Salem. Basic english-hassaniya dictionary. Internet Archive, 2022. URL <https://archive.org/details/eng-hass-dict-a-z>. Public reference containing more than 8,000 words, phrases, and sentences.
- [26] ScriptureEarth. Arabic, hassaniya (mey): Scripture resources. Hassaniya scripture resource index, 2026. URL <https://www.scriptureearth.org/00eng.php?iso=mey>.
- [27] Christos Christodoulopoulos and Mark Steedman. A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation*, 49:375–395, 2015. doi: 10.1007/s10579-014-9287-y. URL <https://christos-c.com/bible/>.
- [28] Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. URL <https://arxiv.org/abs/2312.11805>.
- [29] Hugging Face. Gemma4: Transformers documentation, 2026. URL https://huggingface.co/docs/transformers/model_doc/gemma4. Accessed 12 September 2026.
- [30] Karl Cobbe et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL <https://arxiv.org/abs/2110.14168>.
- [31] Yubo Wang et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024. URL <https://arxiv.org/abs/2406.01574>.
- [32] Mark Chen et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL <https://arxiv.org/abs/2107.03374>.
- [33] Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The Berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL <https://proceedings.mlr.press/v267/patil25a.html>.
- [34] NLLB Team, Marta R. Costa-jussà, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. URL <https://arxiv.org/abs/2207.04672>.
- [35] Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, 2017. URL <https://aclanthology.org/W17-4770/>.
- [36] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*, 2018. URL <https://aclanthology.org/W18-6319/>.
- [37] AlexandriaX-2026 Organizers. AlexandriaX-2026: Context-aware dialectal arabic MT and MT evaluation shared task. Shared task, Subtask 3: Dialectal Arabic MT Error Detection and Classification, 2026. URL <https://alexandriax.dlnlp.ai/>.
- [38] Samar M. Magdy, Fakhraddin Alwajih, Abdellah El Mekki, Wesam El Sayed, and Muhammad Abdul-Mageed. LQM: Linguistically motivated multidimensional quality metrics for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL <https://aclanthology.org/2026.findings-acl.2012/>.