Mining Large Language Models for Low-Resource Language Data: Comparing Elicitation Strategies for Hausa and Fongbe

Mahounan Pericles Adjovi[1], **Roald Eiselen**[2], **Prasenjit Mitra**[1]
[1]Carnegie Mellon University Africa, Kigali, Rwanda
[2]Centre for Text Technology, North-West University, Potchefstroom, South Africa
madjovi@andrew.cmu.edu, Roald.Eiselen@nwu.ac.za, prasenjm@andrew.cmu.edu

Abstract

Large language models (LLMs) are trained on data contributed by low-resource language communities, including curated datasets such as MasakhaNER and MAFAND-MT, yet the linguistic knowledge encoded in these models remains accessible only through commercial APIs. This paper investigates whether strategic prompting can extract usable text data from LLMs for two West African languages: Hausa (Afroasiatic, approximately 80 million speakers) and Fongbe (Niger-Congo, approximately 2 million speakers). We systematically compare six elicitation task types: creative writing, functional text, structured knowledge, dialogue, topic-switching probes, and constrained generation across two commercial LLMs (GPT-4o Mini and Gemini 2.5 Flash). Generated outputs are evaluated on linguistic accuracy, lexical diversity, domain coverage, and code-switching rates through automatic assessment metrics. Our findings reveal that elicitation strategy significantly affects output quality and that optimal strategies differ by language: Hausa benefits from volume-maximizing tasks such as functional text and dialogue, while Fongbe requires constraint-heavy prompts that enforce monolingual output. GPT-4o Mini extracts 6–41$\times$ more usable target-language words per API call than Gemini, though Gemini achieves higher language purity for Fongbe on constrained tasks. We provide a practical framework for low-resource language communities to maximize usable data extraction from LLMs and release all generated corpora and code.

Keywords: low-resource languages, African NLP, data extraction, large language models, Hausa, Fongbe, resource creation

1. Introduction

Natural Language Processing (NLP) technologies remain largely inaccessible to speakers of most African languages due to severe data scarcity (Joshi et al., 2020). Languages such as Fongbe, a national language of Benin, and Hausa, widely spoken across West Africa, suffer from limited digital text resources despite having millions of speakers. Meanwhile, large language models (LLMs) have been trained on web-scale data that includes contributions from these language communities, including curated academic datasets such as MasakhaNER 2.0 (Adelani et al., 2022b) and MAFAND-MT (Adelani et al., 2022a). The linguistic knowledge absorbed from these sources resides within commercial LLMs, yet it flows back to these communities only through paid API access. A natural question arises: can we systematically extract usable language data from these models to create new resources for the very communities whose data helped build them?

This question has both practical and ethical significance. On the practical side, low-resource language communities face a critical bootstrapping problem: building NLP systems requires data, but data collection is expensive and slow. If LLMs can serve as an efficient source of text data across diverse domains, this could accelerate resource creation for languages where text expansion remains a critical priority gap. On the ethical side, the relationship between LLM training data and community benefit is asymmetric: language communities contribute data that increases the commercial value of LLMs, yet receive limited benefit in return. Developing systematic methods for extracting linguistic knowledge from LLMs represents a practical step toward rebalancing this relationship.

Extracting usable language data from LLMs is non-trivial for several reasons. First, LLM generation quality varies dramatically across low-resource languages, with substantial performance gaps documented even among African languages with millions of speakers (Robinson et al., 2023; Hendy et al., 2023). Second, generated text frequently exhibits code-switching with colonial languages: English for Hausa, French for Fongbe, reducing its utility as monolingual training data (Orife et al., 2020). Third, for tonal languages like Fongbe, LLMs frequently produce missing or incorrect diacritics, which are obligatory in standard orthography and distinguish lexical meaning (Lefebvre and Brousseau, 2002). Fourth, it is unclear which prompting strategies maximize both the quantity and quality of extractable data.

Previous work has examined LLMs for low-resource language tasks primarily through machine translation (Robinson et al., 2023; Hendy et al., 2023) or data augmentation for specific downstream tasks (Whitehouse et al., 2023). The Fikira dataset (Adelani et al., 2024) demonstrated that instruction-tuned models can generate reasoning

Proceedings of Resources for African Indigenous Languages (RAIL) 2026 @ LREC 2026, pages 20–30
12 May 2026.

data for African languages, but did not compare across elicitation task types. To our knowledge, no study has systematically explored elicitation strategies to assess which tasks may yield the most usable data per API call for low-resource African languages. This work presents an early exploratory investigation into this question, with the goal of identifying promising directions rather than drawing definitive conclusions.

1.1. Research Question

We address the following central research question:

> *Which types of elicitation tasks maximize the quantity and quality of usable text data that can be extracted from large language models for Hausa and Fongbe?*

This is operationalized through three sub-questions:

1. How does the linguistic quality of LLM-generated text vary across elicitation task types for Hausa and Fongbe?
2. Which elicitation strategies produce the greatest lexical diversity and domain coverage per API call?
3. Do optimal elicitation strategies differ between languages with different levels of LLM support?

1.2. Summary of Contributions

- A systematic taxonomy of LLM elicitation strategies for low-resource language data extraction, evaluated across six task types for two typologically distinct West African languages (Section 3).
- An empirical comparison of two commercial LLMs revealing that GPT-4o Mini generates 6–41 times more usable target-language text than Gemini 2.5 Flash, with language-specific optimal strategies (Section 4).
- A practical framework and released corpora enabling low-resource language communities to replicate our methodology (Section 5).

2. Related Work

2.1. African Language NLP Resources

Research on African language technology has accelerated significantly since 2019. The Masakhane project established a participatory approach to machine translation across more than 30 African languages (Nekoto et al., 2020). Subsequent efforts produced standardized benchmarks: MasakhaNER 2.0 for NER across 20 languages (Adelani et al., 2022b), MasakhaPOS for part-of-speech tagging (Dione et al., 2023), and MAFAND-MT for news-domain machine translation (Adelani et al., 2022a).

Despite these advances, resource availability remains severely unbalanced. Under the taxonomy of Joshi et al. (2020), Hausa falls in mid-tiers (3–4) given its international media presence, while Fongbe falls closer to the lowest tiers (0–1). Continental surveys confirm that most African languages lack sufficient corpora (Hedderich et al., 2021). Our work proposes LLM-based data extraction as a scalable complement to manual corpus construction.

2.2. LLMs for Low-Resource Languages

Robinson et al. (2023) showed that ChatGPT degrades significantly for low-resource African languages. Hendy et al. (2023) found systematic quality drops for languages with limited web presence. AfriDoc-MT (Alabi et al., 2025) evaluated document-level translation for African languages including Hausa. African-centric models such as AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022) outperform multilingual baselines but focus on comprehension rather than generation.

A critical gap persists: none of these studies investigate which *types* of prompts maximize extractable language data. Our work reframes the question from "how well do LLMs translate into language X?" to "which prompting strategies extract the most usable data from LLMs for language X?"

2.3. Data Augmentation and Synthetic Data

Data augmentation encompasses Easy Data Augmentation (EDA) (Wei and Zou, 2019), back-translation (Sennrich et al., 2016), and LLM-based generation (Schick and Schütze, 2021). Whitehouse et al. (2023) found mixed results for low-resource LLM augmentation. Dai and Adel (2020) showed augmentation effectiveness depends on method and dataset size. The Fikira dataset (Adelani et al., 2024) generated reasoning data for African languages but did not compare elicitation strategies. These studies evaluate augmentation for specific downstream tasks rather than investigating which strategies maximize general corpus utility.

3. Methodology

We design a controlled experiment comparing six elicitation task types across two LLMs and two languages. All prompts, scripts, and evaluation code

are released publicly (see Appendix A). Full prompt structures and examples are provided in Appendix D.

3.1. Elicitation Task Taxonomy

Table 1 summarizes six task types, each probing a different dimension of LLM linguistic knowledge (see Appendix D for full prompt details).

Task Type	Rationale	N
Creative Writing (poems, stories, folktales, songs, proverbs)	Tests deep cultural and linguistic knowledge; generates domain-diverse narrative text	25
Functional Text (letters, instructions, news, recipes, announcements)	Tests practical domain coverage; generates text useful for downstream NLP	25
Structured Knowledge (definitions, grammar examples, vocabulary lists, translations)	Tests metalinguistic knowledge; produces high-density lexical output	25
Dialogue (conversations, interviews, negotiations, family discussions)	Tests colloquial register and spoken-form generation	25
Topic Switching (domestic→sports, narrative shifts, knowledge switches)	Tests language maintenance robustness under topic changes	25
Constrained Gen. (vocab-constrained, no-code-switching, technical monolingual)	Tests ability to stay in target language under explicit constraints	25

Table 1: Elicitation task taxonomy: 6 types $\times$ 25 prompts = 150 per language.

Creative Writing prompts request poems, folktales, stories, songs, and proverbs about culturally relevant themes. **Functional Text** prompts request letters, instructions, news articles, recipes, and announcements. **Structured Knowledge** prompts request definitions, cultural explanations, grammar examples, vocabulary lists, and translations. **Dialogue** prompts request conversations in varied social contexts (market, clinic, family, interview, negotiation). **Topic Switching** prompts begin on a familiar topic and switch to a domain typically discussed in colonial languages, requiring continuation in the target language. **Constrained Generation** prompts impose vocabulary constraints, no-code-switching rules, length requirements, and structural formats.

3.2. Languages

Hausa (ISO 639-3: hau) is an Afroasiatic language spoken by approximately 80–100 million people across Nigeria, Niger, and neighboring countries. It features grammatical gender, rich morphology, and complex tense-aspect-mood marking (Newman, 2000). It has a standardized Latin orthography, substantial web presence, and is included in XLM-RoBERTa (Conneau et al., 2020). The colonial contact language is English.

Fongbe (ISO 639-3: fon) is a Niger-Congo Gbe language spoken by approximately 2 million people in Benin. It features serial verb constructions and a three-tone system with obligatory diacritic marking (Lefebvre and Brousseau, 2002). Tone distinguishes meaning: *kó* (high) = "harvest," *kò* (low) = "build," *kô* (falling) = "neck." Fongbe has minimal web presence and is absent from XLM-RoBERTa. The colonial contact language is French.

3.3. Models and Prompt Design

We evaluate GPT-4o Mini (OpenAI) and Gemini 2.5 Flash (Google), both accessed with temperature 0.7, top-p 0.95, max output 1,024 tokens, and a system prompt requiring target-language output. Each prompt template contains placeholders (`{language}`, `{language_culture}`, `{colonial_language}`) substituted at generation time.

All prompts are written in English. This choice reflects a deliberate experimental design decision: English-language prompting provides a controlled, reproducible interface that does not require annotators to be proficient in Hausa or Fongbe, and enables direct comparability across languages. We acknowledge that prompting directly in the target language may yield different results and we consider this a promising direction for future work (Section 6). Preliminary evidence from related work suggests that target-language prompting can improve output quality for well-resourced languages, though its effects for very low-resource languages like Fongbe remain untested.

Each prompt is sent once per model per language: $150 \times 2 \times 2 = 600$ API calls. Outputs are saved as JSON with resumability support.

3.4. Evaluation Framework

We evaluate outputs using: **Output Validity** (minimum 20 tokens); **Lexical Diversity** (TTR, hapax ratio, vocabulary size); **Language Fidelity** via GlotLID (Kargaran et al., 2023), a fastText classifier covering 2,000+ languages including `hau_Latn` and `fon_Latn`, applied at document and sentence levels; **Diacritic Analysis** for Fongbe (tonal vowel ratio); **Repetition Detection** (4-gram and sentence repetition); and **Reference Overlap** (character trigram cosine similarity against MasakhaNER 2.0 training text for Hausa).

4. Results

We report results from 600 API calls (150 prompts $\times$ 2 models $\times$ 2 languages).

4.1. Output Validity

Table 2 reports the percentage of outputs exceeding the 20-token minimum and the average word count per condition.

	Gemini		GPT-4o Mini	
Task	**Fon**	**Hau**	**Fon**	**Hau**
Creative	28/18	76/27	100/90	100/104
Functional	36/17	68/20	100/153	100/205
Structured	40/18	56/20	100/92	100/114
Dialogue	80/23	52/20	100/145	100/190
Topic Switch	4/15	88/73	100/157	100/190
Constrained	20/17	92/34	100/78	100/125
Overall	**35/18**	**72/32**	**100/119**	**100/155**

Table 2: Output validity (% valid/avg. words) by task and model.

GPT-4o Mini achieves 100% validity across all 12 conditions, producing outputs averaging 119 words for Fongbe and 155 for Hausa. Gemini generates much shorter responses: 18 words on average for Fongbe (35% valid) and 32 for Hausa (72% valid). The disparity is most extreme for Fongbe topic-switching, where Gemini produces valid output for only 4% of prompts.

4.2. Language Fidelity

Table 3 reports document-level target language detection using GlotLID.

Hausa outputs are reliably identified: 89% for Gemini and 100% for GPT-4o Mini. Fongbe shows greater variation. Gemini achieves its highest Fongbe fidelity on constrained generation (100%) and topic switching (96%), while GPT-4o Mini scores highest on constrained generation (88%) but poorly on topic switching (32%). When Fongbe outputs are misidentified, GlotLID most frequently labels them as English (32 cases), Yoruba (24), or French (23), suggesting code-switching contamination or generation in related Gbe languages.

	Gemini		GPT-4o Mini	
Task	**Fon**	**Hau**	**Fon**	**Hau**
Creative	60	92	48	100
Functional	68	96	40	100
Structured	40	72	52	100
Dialogue	12	76	60	100
Topic Switch	96	100	32	100
Constrained	100	100	88	100
Overall	**63**	**89**	**53**	**100**

Table 3: Document-level target language detection (%) by GlotLID, per task and model.

At the sentence level, constrained generation achieves the lowest code-switching rates across both models (0.01–0.09). GPT-4o Mini shows consistently low code-switching for Hausa (0.01–0.16) but elevated rates for Fongbe (0.25–0.66), indicating frequent interspersion of French or English sentences within otherwise Fongbe text.

4.3. Lexical Diversity

Table 4 reports TTR and vocabulary size.

	Gemini		GPT-4o Mini	
Task	**Fon**	**Hau**	**Fon**	**Hau**
Type-Token Ratio				
Creative	.93	.92	.58	.67
Functional	.88	.92	.48	.60
Structured	.96	.95	.60	.71
Dialogue	.88	.92	.46	.58
Topic Switch	.89	.81	.48	.63
Constrained	.89	.82	.54	.67
Avg. Vocabulary Size				
Creative	16	24	50	68
Functional	15	19	74	117
Structured	18	19	48	73
Dialogue	20	18	66	108
Topic Switch	13	53	70	117
Constrained	15	27	32	74

Table 4: Lexical diversity measured by Type-Token Ratio (TTR) and average vocabulary size per condition.

Gemini's higher TTR (0.81–0.96 vs. 0.46–0.71) is largely an artifact of output length. In absolute terms, GPT-4o Mini yields 13,895 unique Hausa and 8,478 Fongbe word tokens across all outputs, versus Gemini's 3,977 and 2,427—a 3.5$\times$ advantage.

4.4. Fongbe Diacritic Analysis

GPT-4o Mini produces diacritics in 96–100% of Fongbe outputs, with diacritic-to-alphabetic ratios of 0.24–0.37. Gemini is less reliable: only 36% of dialogue outputs contain diacritics (ratio 0.02), versus 100% for constrained generation (ratio 0.31). Explicit constraints help Gemini activate Fongbe orthographic knowledge that unconstrained tasks fail to elicit.

4.5. Extraction Efficiency

Table 5 presents usable words per API call (from outputs that are both valid and detected as the target language). Figure 1 (Appendix B) visualizes these differences.

Model	Task	Fon	Hau
Gemini	Creative	0.0	20.7
	Functional	1.7	13.5
	Structured	0.9	5.8
	Dialogue	0.0	7.0
	Topic Switch	0.0	70.6
	Constrained	5.7	32.7
	Per call	**1.4**	**25.0**
GPT-4o	Creative	37.5	103.7
	Functional	55.0	204.8
	Structured	49.4	114.5
	Dialogue	80.8	190.0
	Topic Switch	50.6	190.1
	Constrained	69.6	124.6
	Per call	**57.2**	**154.6**

Table 5: Usable target-language words per API call. GPT-4o Mini is 6× more efficient for Hausa, 41× for Fongbe.

GPT-4o Mini extracts 154.6 usable Hausa words per call (6× Gemini) and 57.2 Fongbe words (41× Gemini). The most efficient strategies differ by language: for Hausa, functional text and dialogue maximize extraction (190–205 words/call); for Fongbe, dialogue (80.8) and constrained generation (69.6) are most productive. Gemini's Fongbe extraction is near zero for most tasks.

Repetition rates are low across all conditions ($<$0.06). Reference corpus overlap shows GPT-4o Mini's Hausa outputs have higher character trigram similarity to MasakhaNER 2.0 text (cosine 0.10 vs. 0.07 for Gemini). Crucially, both values are well below 0.15, a conservative threshold above which near-verbatim reproduction would become plausible. The elevated similarity for GPT-4o Mini most likely reflects that this model generates more natural Hausa text whose statistical profile resembles existing Hausa corpora—an indicator of generation quality rather than memorization. All cosine values by task type are visualized in Figure 4 (Appendix B).

5. Discussion

5.1. Optimal Elicitation Strategies by Language

Our results confirm that optimal strategies differ substantially between languages (RQ3).

For **Hausa**, functional text and dialogue yield the most usable words (190–205 per call with GPT-4o Mini), while constrained generation and topic switching achieve the highest language fidelity (100% for both models). Hausa is sufficiently represented in LLM training data to sustain generation across diverse task types.

For **Fongbe**, constrained generation emerges as the most reliable strategy: highest language fidelity (100% Gemini, 88% GPT-4o Mini), best diacritic ratios, and lowest code-switching. Communities working with extremely low-resource languages should prioritize constrained generation prompts that explicitly require monolingual output and specify target-language vocabulary.

The Hausa–Fongbe gap is consistently large: GPT-4o Mini achieves 100% vs. 53% language fidelity, produces 2.7× more usable words per call, and exhibits 4–10× lower code-switching rates. This disparity likely reflects training data representation rather than inherent linguistic difficulty.

5.2. Training Data as a Confounding Factor

The performance gap between GPT-4o Mini and Gemini 2.5 Flash and between Hausa and Fongbe most plausibly reflects differences in training data composition rather than architectural differences per se. Robinson et al. (2023) demonstrated that ChatGPT performance degrades sharply for languages underrepresented in web-crawled pretraining data, and Hendy et al. (2023) showed systematic quality drops correlate with web presence rather than linguistic complexity. Hausa has a substantial international media presence (BBC Hausa, VOA Hausa), whereas Fongbe has minimal digital footprint. If Gemini's training data includes proportionally less Hausa and Fongbe text than GPT-4o Mini's, this would fully explain the extraction efficiency gap without invoking any architectural cause. Unfortunately, neither OpenAI nor Google discloses the language-level composition of their training data, making this hypothesis untestable with current information. Future work using open-weight models with documented training corpora (e.g., BLOOM, Llama variants) could help disentangle data from architecture effects.

5.3. Implications for Resource Creation

Our findings yield practical recommendations. First, **model selection matters more than task selection**: switching from Gemini to GPT-4o Mini increases Fongbe efficiency by 41×, whereas task variation within GPT-4o Mini yields only 2× difference. Second, **explicit constraints improve fidelity**: constrained generation consistently achieves the highest language purity and diacritic accuracy. Third, **post-hoc filtering is essential**: even the best Fongbe condition produces 12% non-target outputs; GlotLID filtering can remove contaminated text. Fourth, **cost-efficiency is compelling**: GPT-4o Mini extracted 23,192 usable Hausa words and 8,574 Fongbe words for under $0.10, scalable to substantial corpora for under $10.

6. Conclusion and Future Work

We presented an exploratory evaluation of six LLM elicitation strategies for extracting usable text data for Hausa and Fongbe. While the scale of this study is limited, our initial findings suggest three trends worth investigating further. First, GPT-4o Mini produces substantially more usable text than Gemini 2.5 Flash, yielding 6× more Hausa words and 41× more Fongbe words per API call. Second, elicitation strategies appear to be language-dependent: Hausa benefits from volume-maximizing tasks (functional text, dialogue), while Fongbe appears to require constraint-heavy prompts (constrained generation). Third, the Hausa–Fongbe performance gap is consistent across conditions, suggesting that LLM-based extraction may currently be more viable for moderately resourced languages. These findings are preliminary and will require validation at larger scale and across additional languages and models.

Future work will include additional LLMs (Claude Sonnet, open-source and African-language-focused models), human evaluation with native speakers, downstream utility testing on MasakhaNER 2.0 and MasakhaPOS, target-language prompting experiments, larger prompt samples for statistical robustness, and extension to additional African languages.

7. Limitations

This study has several limitations. First, we evaluate only two commercial LLMs; the performance gap we observe may not generalize to open-source or African-language-focused models. Second, our evaluation relies entirely on automatic metrics; human evaluation by native speakers is essential, particularly for Fongbe where GlotLID misidentifies 47% of GPT-4o Mini outputs despite many likely containing valid Fongbe with code-switching — misidentification does not necessarily imply low linguistic quality, but may reflect code-switching that the classifier penalizes. Third, we do not evaluate downstream task utility: whether extracted corpora improve NER or POS tagging performance remains to be tested. Fourth, with 25 prompts per task type, sample sizes are modest; while directional patterns are consistent across conditions, larger experiments would enable more robust statistical comparisons. Fifth, all prompts are written in English; target-language prompting may yield different results. Sixth, our memorization assessment relies on reference overlap as an indirect proxy, though the uniformly low cosine similarity values ($<$0.12) suggest verbatim reproduction is unlikely. Finally, our methodology relies on commercial APIs, introducing cost barriers and reproducibility concerns; future work should investigate open-source alternatives.

8. Ethics Statement

Data provenance and community benefit. We acknowledge that LLMs were trained on data contributed by language communities, often without explicit consent. Our work aims to redirect encoded knowledge back to these communities. All generated data will be released under CC-BY-4.0.

Quality and potential harms. LLM-generated text may contain errors, inaccuracies, or hallucinated content. We document the synthetic nature of all corpora and recommend native speaker validation before production use.

Commercial API usage. Our methodology relies on commercial APIs, introducing cost barriers. Future work should investigate open-source alternatives.

9. Data and Code Availability

All generated corpora, prompts, generation scripts, and evaluation code will be made publicly available upon acceptance under a CC-BY-4.0 license. The evaluation pipeline, including the GlotLID-based language fidelity assessment, is provided for full reproducibility.

10. Acknowledgements

The authors thank the reviewers for their constructive feedback. We acknowledge the Masakhane community for their foundational contributions to African NLP resources, particularly the MasakhaNER 2.0 and MasakhaPOS datasets used in our evaluation. This publication was developed as part of the Center for Inclusive Digital Transformation of Africa (CIDTA) and the

Afretec Network, which is managed by Carnegie Mellon University Africa and receives financial support from the Mastercard Foundation. The views expressed in this document are solely those of the authors and do not necessarily reflect those of Carnegie Mellon University or the Mastercard Foundation.

11. Bibliographical References

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, et al. 2022a. A few thousand translations go a long way! Leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Adelani, Shamsuddeen Muhammad, et al. 2024. Fikira: Multilingual reasoning dataset for African languages. Masakhane Project Technical Report.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh Dione, et al. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jesujoba Alabi, David Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Jesujoba Oluwadara Alabi, Israel Abebe Azime, et al. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Cheikh M Bamba Dione, David Adelani, et al. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Michael Hedderich, Lukas Lange, Heike Adel, Jannik Strobe, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Rauber, et al. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Pratik Joshi, Sebastin Santy, Amar Buber, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. https://huggingface.co/cis-lmu/glotlid. Version 1.0.

Claire Lefebvre and Anne-Marie Brousseau. 2002. *A Grammar of Fongbe*. Mouton de Gruyter, Berlin.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, et al. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

Paul Newman. 2000. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings*

of the 1st Workshop on Multilingual Representation Learning. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Bonaventure Dossou, Chris Emezue, et al. 2020. Masakhane – machine translation for Africa. *arXiv preprint arXiv:2003.11529*.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Chenxi Whitehouse et al. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Appendix A. Code and Data Repository

All prompts, generation scripts, evaluation code, and generated corpora are publicly available at:
https://github.com/Pericles001/mining_llm_low_resource_languages_fon_hau/tree/main
The repository is organised as follows:

`prompts/` JSON files containing all 150 prompts per language, organised by task type

`src/` Core modules for generation (`generator.py`), evaluation (`evaluator.py`), and language detection (`language_detector.py`)

`scripts/` CLI entry points for generation, evaluation, and analysis

`outputs/` Raw LLM outputs organised by model, language, and task type

`results/` Aggregated evaluation results, figures, and LaTeX tables

Appendix B. Supplementary Figures

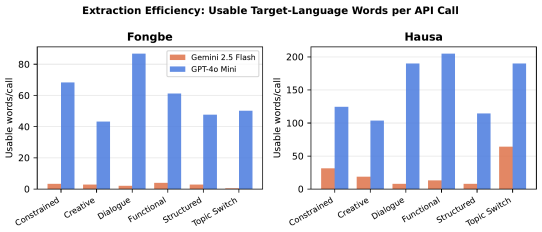

Figure 1: Extraction efficiency: usable target-language words per API call, by model, language, and task type. GPT-4o Mini dominates across all conditions; the gap is most extreme for Fongbe.

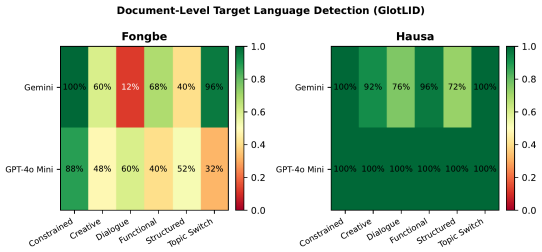

Figure 2: Document-level target language detection heatmap (GlotLID). Green = high fidelity; red = low. Hausa is uniformly high for GPT-4o Mini; Fongbe fidelity depends strongly on task type.

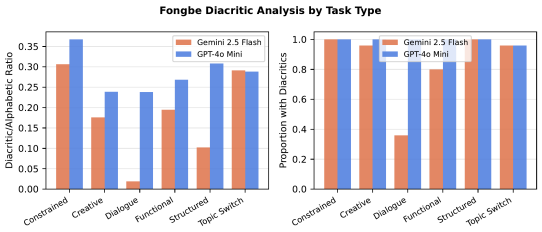

Figure 3: Fongbe diacritic analysis by task type. Left: diacritic-to-alphabetic ratio; Right: proportion of outputs containing any diacritics. Constrained generation reliably elicits diacritics from both models.

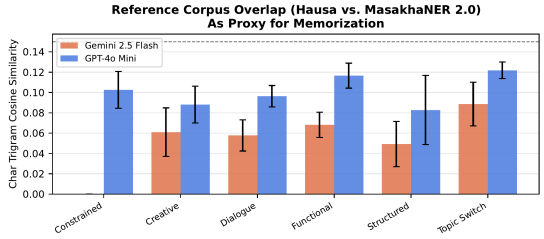

Figure 4: Character trigram cosine similarity between generated Hausa text and MasakhaNER 2.0 training text, used as a proxy for potential memorization. All values are well below 0.15 (dashed line), suggesting outputs represent novel generation rather than training data reproduction.

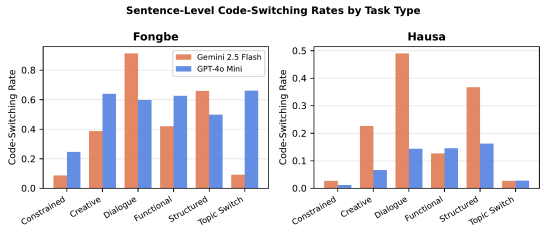

Figure 5: Sentence-level code-switching rates by model, language, and task type. Constrained generation consistently achieves the lowest code-switching. Fongbe shows much higher rates than Hausa across all tasks.

Appendix C. Full Evaluation Summary

Table 6 reports all evaluation metrics across all 24 conditions (2 models $\times$ 2 languages $\times$ 6 task types). *Quality* is a composite score averaging language confidence and inverse code-switching rate.

Model	Lang	Task	Valid%	Words	TTR	Hapax	Vocab	CS	LangConf	Quality
Gemini	fon	constrained	0.20	17.1	0.891	0.802	14.7	0.087	0.998	0.927
		creative	0.28	17.6	0.932	0.876	16.4	0.387	0.782	0.791
		dialogue	0.80	22.8	0.880	0.787	20.2	0.913	0.744	0.555
		functional	0.36	16.7	0.883	0.795	14.7	0.420	0.929	0.816
		structured	0.40	18.5	0.955	0.915	17.7	0.660	0.616	0.645
		topic switch	0.04	15.0	0.892	0.800	13.4	0.093	0.995	0.941
	hau	constrained	0.92	34.0	0.822	0.704	27.2	0.027	1.000	0.921
		creative	0.76	26.8	0.918	0.854	23.5	0.227	0.912	0.867
		dialogue	0.52	19.9	0.920	0.859	18.2	0.490	0.873	0.817
		functional	0.68	20.1	0.923	0.853	18.5	0.127	0.995	0.914
		structured	0.56	20.1	0.946	0.904	18.9	0.367	0.856	0.779
		topic switch	0.88	72.8	0.812	0.707	52.7	0.027	1.000	0.919
GPT-4o	fon	constrained	1.00	77.6	0.544	0.375	31.7	0.246	0.937	0.869
		creative	1.00	90.0	0.581	0.405	50.0	0.640	0.703	0.688
		dialogue	1.00	144.5	0.458	0.275	65.5	0.598	0.868	0.736
		functional	1.00	153.0	0.479	0.325	73.5	0.627	0.870	0.675
		structured	1.00	91.8	0.597	0.486	48.1	0.498	0.862	0.731
		topic switch	1.00	156.8	0.477	0.321	70.4	0.661	0.828	0.646
	hau	constrained	1.00	124.6	0.667	0.520	73.6	0.012	1.000	0.898
		creative	1.00	103.7	0.674	0.512	67.5	0.066	1.000	0.887
		dialogue	1.00	190.0	0.578	0.408	107.6	0.144	1.000	0.871
		functional	1.00	204.8	0.602	0.448	117.4	0.146	1.000	0.874
		structured	1.00	114.5	0.708	0.603	73.1	0.163	0.930	0.889
		topic switch	1.00	190.1	0.628	0.489	116.6	0.028	1.000	0.891

Table 6: Full evaluation summary across all 24 conditions. CS = code-switching rate; LangConf = GlotLID language confidence score; Quality = composite score.

Appendix D. Prompt Taxonomy Details

This appendix documents the structure and rationale of all 150 prompts per language (6 task types $\times$ 25 prompts). Each task type is divided into subtasks to ensure domain coverage. All prompts use three placeholders: `{language}`, `{language_culture}`, and `{colonial_language}`, substituted at generation time.

A. Constrained Generation (cg_01–cg_25)

Subtasks: *vocabulary-constrained* (cg_01–05), *no-code-switching* (cg_06–10), *length-constrained* (cg_11–15), *technical-monolingual* (cg_16–20), *structure-constrained* (cg_21–25).
Design rationale: Constrained generation prompts impose explicit linguistic constraints to prevent code-switching and test the model's ability to generate monolingual output. Vocabulary-constrained prompts seed the output with target-language words, reducing the risk of the model falling back to colonial language vocabulary for unknown concepts. Technical-monolingual prompts specifically target domains (computing, electricity, banking) where Fongbe and Hausa lack standard terminology, forcing the model to paraphrase rather than borrow.
Representative templates:

- `cg_01`: "Write a short paragraph in {language} using ALL of the following words: {word_list_1}. Do not use any {colonial_language} words."
- `cg_06`: "Write a story in {language} about a day at the market. You must write ONLY in {language}. If you do not know a word in {language}, describe the concept using other {language} words instead of switching to {colonial_language}."

Word lists for vocabulary-constrained prompts are provided in the released data.

B. Creative Writing (cw_01–cw_25)

Subtasks: *poem* (cw_01–05), *folktale* (cw_06–10), *story* (cw_11–15), *song* (cw_16–20), *proverb* (cw_21–25).
Design rationale: Creative writing prompts test deep cultural and linguistic knowledge by eliciting culturally rooted content (folktales, proverbs) that requires the model to draw on language-specific cultural knowledge, not just translation of English concepts. Folktales and proverbs are particularly valuable as they are community-specific and cannot easily be produced by back-translation.
Representative templates:

- `cw_06`: "Write a traditional folktale in {language} about a clever tortoise who outsmarts a lion. The story should be 5–10 sentences long."

C. Dialogue (dl_01–dl_25)

Subtasks: *conversation* (dl_01–05), *professional* (dl_06–10), *family* (dl_11–15), *interview* (dl_16–20), *negotiation* (dl_21–25).
Design rationale: Dialogue prompts elicit colloquial register and spoken-form text, which is underrepresented in formal corpora. The negotiation and professional subtasks target domains with specialized vocabulary (medical, agricultural, financial), which helps expand domain coverage of the resulting corpus.

D. Functional Text (ft_01–ft_25)

Subtasks: *letter* (ft_01–05), *instructions* (ft_06–10), *news* (ft_11–15), *recipe* (ft_16–20), *announcement* (ft_21–25).
Design rationale: Functional text prompts target practical domains that are immediately useful for downstream NLP tasks (e.g., news classification, instruction following). These genres are typically well-represented in NLP benchmarks but underresourced for African languages.

E. Structured Knowledge (sk_01–sk_25)

Subtasks: *definition* (sk_01–05), *cultural explanation* (sk_06–10), *grammar examples* (sk_11–15), *vocabulary list* (sk_16–20), *translation* (sk_21–25).
Design rationale: Structured knowledge prompts elicit the model's metalinguistic knowledge, producing high-density lexical output (vocabulary lists, grammar examples) that is directly usable for dictionary construction and grammar documentation.

F. Topic Switching (ts_01–ts_25)

Subtasks: *domestic-to-sports* and related binary switches (ts_01–10), *narrative shift* (ts_11–15), *multi-topic* (ts_16–20), *knowledge switch* (ts_21–25).
Design rationale: Topic-switching prompts stress-test language maintenance by requiring the model to continue in the target language after transitioning to a domain (technology, politics, science) that is more commonly discussed in the colonial contact language. This probes whether language fidelity holds under topic-induced pressure to code-switch.
Representative templates:

- `ts_25`: "In {language}, describe a funeral ceremony. Then, in the same response and still in {language}, explain what artificial intelligence is."