AmharicStoryQA: A Multicultural Story Question Answering Benchmark in Amharic

Israel Abebe Azime[1], Abenezer Kebede Angamo[2], Hana Mekonen Tamiru[3], Dagnachew Mekonnen Marilign[4], Philipp Slusallek[1], Seid Muhie Yimam[5], Dietrich Klakow[1]

[1] Saarland University, [2] AIMS AMMI, [3] Resonance AI4D Lab, Addis Ababa University,
[4] HILCOE School of Computer Science and Technology, [5]University of Hamburg

Abstract

With the growing emphasis on multilingual and cultural evaluation benchmarks for large language models, language and culture are often treated as synonymous, and performance is commonly used as a proxy for a model's understanding of a given language. In this work, we argue that such evaluations overlook meaningful cultural variation that exists within a single language. We address this gap by focusing on narratives from different regions of Ethiopia and demonstrate that, despite shared linguistic characteristics, region-specific and domain-specific content substantially influences language evaluation outcomes. To this end, we introduce ***AmharicStoryQA***, a long-form story question answering benchmark grounded in culturally diverse narratives from Amharic-speaking regions. Using this benchmark, we reveal a significant narrative understanding gap in existing LLMs, highlight pronounced regional differences in evaluation results, and show that supervised fine-tuning yields uneven improvements across regions and evaluation settings. Our findings emphasize the need for culturally grounded benchmarks that go beyond language-level evaluation to more accurately assess and improve narrative understanding in low-resource languages.

1 Introduction

Evaluating Large Language Models (LLMs) on low-resource languages is essential for uncovering their true language understanding capabilities in specific linguistic and cultural contexts, especially given that these models are predominantly trained on large-scale English corpora.

Despite increasing efforts to include more languages in LLM evaluations, particularly African-centric languages, there remains a notable gap in the creation of multicultural evaluation datasets that capture the diversity of linguistic and cultural contexts (Alabi et al., 2025b). Another important limitation in low-resource LLM evaluation is the scarcity of generation-focused datasets designed to simulate real-life model usage, which are essential for ensuring that benchmarking accurately reflects how LLMs perform in practical, everyday scenarios.

To capture the diversity of linguistic and cultural contexts, most recent works (Havaldar et al., 2023; Azime et al., 2025b) expand evaluations by adding more languages and incorporating culturally grounded information drawn from the regions they represent. This idea, *which treats each language as being associated with a single cultural identity, has a solid foundation but still fails to capture the cultural variations that exist within the same language community*. These variations arise from differences in governmental structures, regional histories, social norms, migration patterns, and intra-country ethnic diversity. As a result, evaluations risk oversimplifying cultural nuance and underrepresenting the lived experiences of multilingual and multicultural societies

In this work, we introduce ***AmharicStoryQA***, a multicultural story-based question answering benchmark available in Amharic and English. The benchmark is constructed from pre-translated stories and folktales collected from nine different regions of Ethiopia. Our goal is to provide an evaluation resource that captures Ethiopia's rich regional diversity while enabling the study of long-form question answering. *AmharicStoryQA* focuses on culturally grounded narrative understanding and offers one of the first long-context QA benchmarks centered on Ethiopian stories across nine regions.

Based on the benchmark we introduce, we aim to answer the following research questions: (**RQ1**) *How well do Large Language Models comprehend Amharic narratives?* This is measured through performance on a story-based question answering datasets. (**RQ2**) *How cultural differences influence LLM performance in a single language?* and

Proceedings of the 10th Widening NLP Workshop, pages 20–36
October 28, 2026 ©2026 Association for Computational Linguistics

(**RQ3**) *Can we improve LLM performance using the culturally grounded training dataset we introduce, demonstrating the importance of such data for addressing the limitations observed?* To support these questions, we further conduct human error analysis to better identify the linguistic and cultural factors contributing to model failures.

To address this question, our work makes the following contributions.

- We introduce ***AmharicStoryQA***, a multicultural story-based question answering benchmark in Amharic and English, consisting of 571 training and 649 test examples derived from 244 stories collected across nine Ethiopian regions. The training and test sets are constructed with disjoint stories, ensuring that no story appears in both splits and thereby minimizing the risk of story-level data leakage. The benchmark includes both multiple-choice question answering (MCQA) and open-ended generation questions, enabling comprehensive evaluation of model understanding across formats. [1]

- We explore the zero-shot performance of seven opensource LLMs and analyze how their results vary across regions, languages, prompt languages, and task types with human error analysis.

- We explore the effect of supervised fine-tuning on improving the performance of selected LLMs.

2 Related Work

Low Resource LLM Evaluations. Low-resource conditions, characterized by the lack of datasets for specific tasks and domains, limited digitization of languages, and insufficient data to effectively train and improve models, are common across many African languages (Nigatu et al., 2024). Evaluating Large Language Models (LLMs) in such settings presents unique challenges, primarily due to data scarcity and the dominance of English-centric benchmarks (Singh et al., 2025; Anugraha et al., 2026). Recent efforts have begun to address this gap by introducing multilingual evaluation benchmarks that include African languages (Adelani et al., 2025; Ojo et al., 2025). Despite the great deal of work done to include number of tasks and languages and cultural values (Romero et al., 2024; Myung et al., 2024) there is still a need to deeper exploration in single language multiple cultural evaluations.

Long Sequence LLM Evaluations. While the capability of LLMs to process long contexts is a rapidly evolving area of research, benchmarks evaluating this capability are predominantly available only for high-resource languages. Recent works such as Alabi et al. (2025a) evaluate long-context document translation across five low-resource languages, including Amharic. Despite these efforts, there remains a critical absence of long-form benchmarks for Amharic that evaluate an LLM's ability to maintain coherence and track narrative arcs over extended contexts, leaving a significant blind spot in understanding how modern LLMs handle extensive texts in Ethiopian languages.

Story Question Answering. Story-based question answering assesses the ability of the model to comprehend narrative structures, character motivations, and causal chains. AfriQA (Ogundepo et al., 2023) introduces a cross-lingual open-retrieval QA dataset covering seven African languages, evaluating models on their ability to retrieve and answer questions using evidence that is often available only in English or French. Similarly, Belebele (Bandarkar et al., 2024) offers a parallel reading comprehension benchmark across 122 language variants, including Amharic. However, these evaluations typically rely on short passages or translated content, which may not fully capture the linguistic depth required for native-level understanding. Furthermore, multilingual benchmarks often treat languages such as Amharic as monolithic, overlooking the need for region-specific evaluations that account for local variations in dialect and usage.

Cultural Evaluations Truly robust LLMs must capture not only the syntactic structure of a language but also its cultural context and dialectal variation. Recent work has examined the fairness and robustness of LLMs across different dialects of English, even when a single language is used as the communication medium (Srirag et al., 2025). Other studies leverage culturally grounded proverbs and sayings across multiple languages to more rigorously evaluate language understanding capabilities (Cecilia Liu et al., 2024; Azime

[1] https://huggingface.co/collections/israel/amharicstoryqa

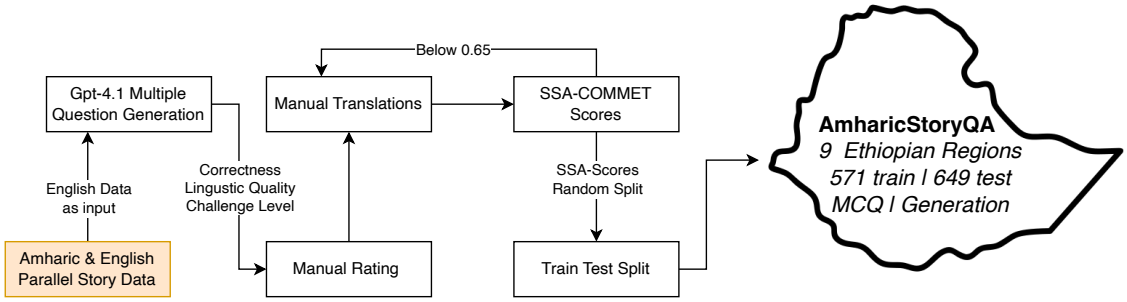


Figure 1: **Overview of the *AmharicStoryQA* dataset** creation process, from Amharic and English parallel stories through question generation, manual translation rating, quality control, and train–test split verification.

et al., 2025b). In this work, we adopt a single-language, multi-cultural evaluation framework and extend prior efforts by introducing long-form, culturally grounded, story-based question answering, enabling a more comprehensive assessment of language understanding across cultural variations.

Amharic Question Answering Recent work such as (Taffa et al., 2024; Belay et al., 2022) introduces an extensive Amharic question answering benchmark used to evaluate Amharic QA capabilities; however, it focuses primarily on open-ended question answering, which is more challenging to evaluate reliably. In this work, we contribute a multi-task story-based question answering dataset that spans multiple cultures in same language setting, resulting in broader cultural diversity and improved regional coverage additional to narrative style long sequence source documents that was not included in previous works.

3 Methodology

3.1 Dataset

The stories used in this dataset are sourced from Ethiopian folktales[2] webpages which was previously included for story generation tasks in low resource LLM work Azime et al. (2024).

As shown in Figure 1, we followed the steps described below, building on prior human-in-the-loop dataset creation approaches (Yuen et al., 2025; Berkane et al., 2025).

We worked with 244 stories collected from nine distinct regions, with the SNNP region further subdivided into multiple zonal classifications, as detailed in Appendix B. We will release the dataset strictly for research purposes only. Figure 3 illustrates the *Llama 3.1 8B Instruct*token counts (sequence lengths) for each story across regions showing coverage of large sequence stories for this work. Average sequence length of stories by region is shown in Table 3.

3.1.1 Question Generation

Given that we have verified English question answer pairs for all Amharic stories collected and presented in Ethiopian folktales webpages, we leveraged Gpt-4.1 to generate batches of five questions per story. This approach enabled us to produce five distinct questions for each story without repeating ideas across them. For each generated question, we also produced one correct answer and three distractor choices: (choice B) a choice biased by world knowledge, (Choice C) an unrelated statement, and (Choice D) a factually false option. We generated 5 questions from each story in parallel to avoid repetitive questions. All content was generated in a single step using structured generation [3], which helped prevent parsing inconsistencies that can arise when handling multiple separate generations.

3.1.2 Manual Question Rating

After obtaining the English stories, Amharic stories, four distractor choices, and the correct answer, we developed a question-rating guideline to ensure that the generated English questions met standards of correctness and faithfulness, linguistic quality and clarity, and comprehension depth and challenge level. The evaluation was conducted by three human annotators. **Correctness and Faithfulness** evaluates whether the question is grounded in the

[2]https://www.ethiopianfolktales.com/am

[3]https://platform.openai.com/docs/guides/structured-outputs

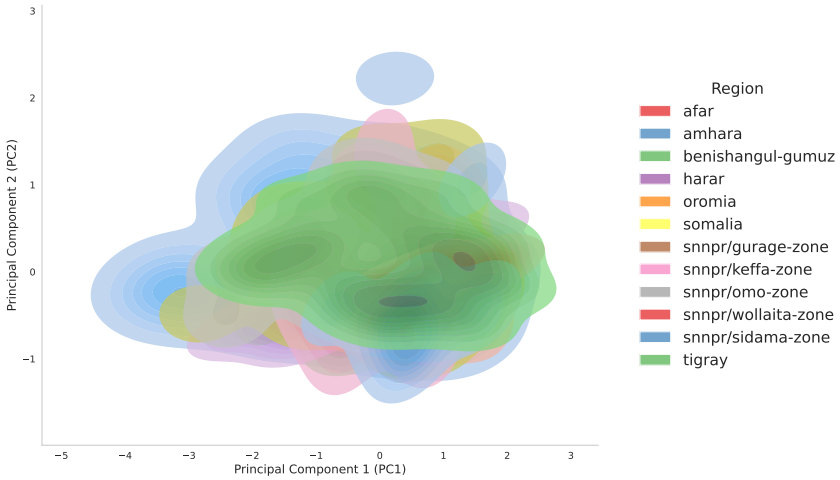


Figure 2: **Regional semantic density of Amharic stories across regions** in the *AmharicStoryQA* corpus. Each contour shows a kernel density estimate (KDE) over 2D PCA projections of multilingual MiniLM sentence embeddings, coloured by region. The plot shows that regional stories largely share a common semantic space, with subtle outer-region contours revealing culturally specific variations within the same language.

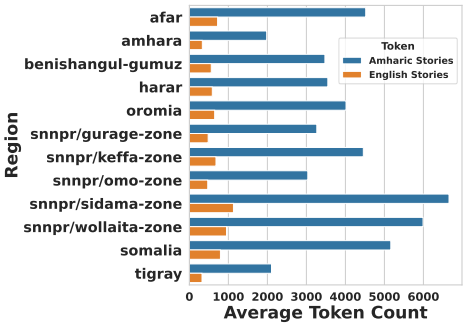


Figure 3: Average Stories Sequence Length by Region

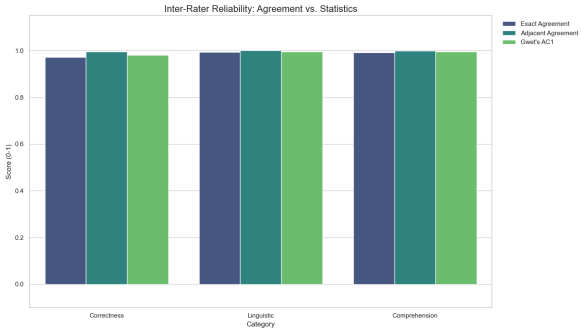


Figure 4: Inter-Rater Reliability agreement in Correctness, Linguistic, and Comprehension categories assessed with three human evaluators

story, whether the correct answer is fully supported by the narrative, and whether the distractors follow the intended structure (plausible, world-knowledge-biased, irrelevant, or factually false). **Linguistic Quality and Clarity** assesses whether the question is grammatically correct, fluent, and natural, particularly for Amharic, and whether it is easy to understand and translate. **Comprehension Depth and Challenge Level** examines whether the question tests meaningful understanding of the story, including inferential or reasoning-based comprehension rather than simple surface-level recall. We used Gwet's AC1 to measure agreement between raters beyond chance, as it is more robust to class imbalance and prevalence effects. Exact annotation guideline can be found in Appendix F.

3.1.3 Manual Translations

We translated each question and its answer choices into Amharic using experienced annotators who had previously worked on translation tasks. The translators were familiar with the narrative content, enabling them to preserve meaning, style, and cultural relevance across all generated questions.

3.1.4 SSA-COMET scoring

To evaluate machine translation (MT) quality for our low resource African language setting, we adopt both reference-based and reference-free evaluation metrics introduced in Li et al. (2025). These metrics allow us to systematically assess the quality of the human-translated Amharic outputs. Following prior work (Alabi et al., 2025a; Azime et al., 2025a), we flagged all translations with a score below 0.65 for manual inspection. While we found that many low-scoring cases were in fact false negatives produced by SSA-COMET (Li et al., 2025), reflecting known limitations of automatic evaluation in low-resource languages, we nonetheless identified and corrected a small number of gen-

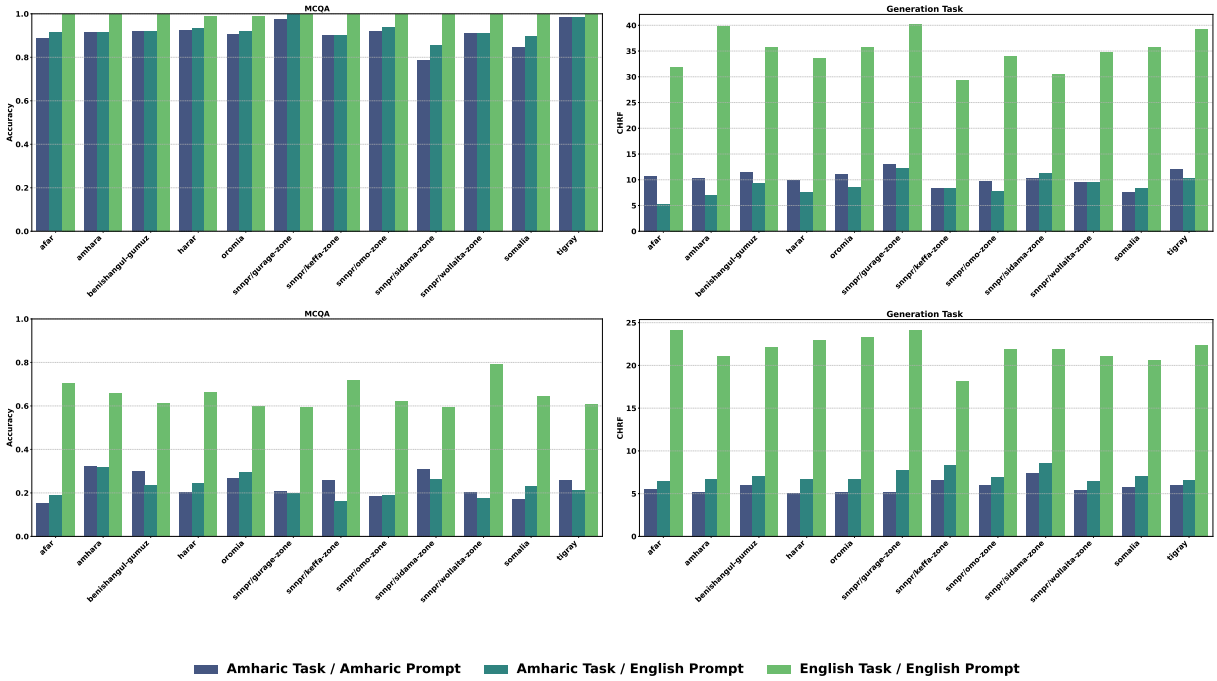


Figure 5: **Regional Performance.** Performance of the two differently performing models (top) (*Gemma-3-27b-it*) and (*Llama 3.2 1B Instruct*), showing performance variation across narratives from different regions in our dataset.

uine translation errors attributable to the translators. This combined automatic-plus-manual verification ensured that our final Amharic translations maintained high linguistic quality and fidelity to the original stories.

3.1.5 Train-Test Split

We sorted all story and question–answer pairs by their average SSA-COMET scores and used this ordering to create the train–test split. This approach ensures that translation inconsistencies do not disproportionately influence model evaluation and failures due to translation problems does not result evaluation problem.

After randomizing the answer choices from the structured-generation labels (where the correct answer was always originally assigned to option A) to simplify downstream quality control, we created a balanced 50/50 split for the benchmark. This split was lightly informed by the overall SSA-COMET score distribution to avoid translation-quality imbalance, resulting in 571 training instances and 649 test instances. This include instances where ssa-comet fails to capture translation quality according to human raters.

4 Experimental Setup

4.1 Evaluation

We selected **7** models that support a maximum context length of 128k tokens to align with the long-form nature of our tasks.

We conducted all evaluations using EleutherAI's open-source Language Model Evaluation Harness (`lm-eval`) (Gao et al., 2024). The framework supports multiple evaluation strategies, including log-likelihood, text generation, and perplexity, with experiments configured and managed via configuration files. For open-source models, we employed log-likelihood and generation-based evaluation for multiple-choice tasks. In multiple-choice each candidate option is appended to the corresponding question prompt, after which the log-likelihood is computed. Model accuracy is then reported based on the option with the highest log-likelihood score.

4.2 Finetuning

To fine-tune LLLMs, we leveraged **LLaMA-Factory** (Zheng et al., 2024), an open-source framework that supports efficient fine-tuning through parameter-efficient methods and scalable training pipelines. We performed LoRA-based supervised fine-tuning using two A100 GPUs on English-only, Amharic-only, and combined multilingual datasets. All datasets were converted to the

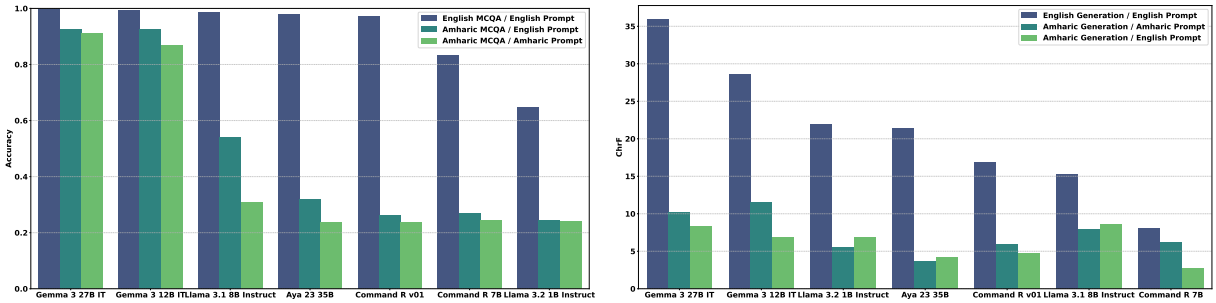


Figure 6: **Results on Amharic vs. English tasks.** Left: Accuracy on multiple-choice question answering (MCQA) tasks under different task–prompt language settings. Right: Generative performance measured using chrF under the same conditions.

Alpaca instruction format and will be shared as part of our released resources. Training was conducted for three epochs using a LoRA rank of 8 and an effective batch size of 8 via gradient accumulation. A systematic exploration and optimization of hyperparameters, including LoRA configuration and training dynamics, is left for future work.

4.3 Task choice

Previous work has shown that large language models' performance on multiple-choice question answering (MCQA) tasks is heavily influenced by factors such as option ordering and other structural biases (Loginova et al., 2025). These limitations are further amplified in low-resource language settings, as demonstrated by recent analyses in (Azime et al., 2025b). Despite these known issues, MCQA remains one of the most widely used evaluation paradigms due to its ease of benchmarking, reproducibility, and straightforward metric computation. For this reason, we include MCQA tasks in our evaluation while complementing them with generative tasks, allowing for a more balanced assessment across different task formulations and reducing reliance on a single evaluation signal.

5 Result and Analysis

Cultural Differences Following Cecilia Liu et al. (2024), we visualize regional semantic variation using kernel density estimates (KDE) over 2D PCA projections of multilingual MiniLM sentence embeddings. As shown in Figure 2, the substantial overlap among regions suggests that stories written in Amharic largely occupy a shared semantic space, reflecting common narrative structures, themes, and stylistic conventions across Ethiopia. At the same time, the outer contour differences observed for specific regions reveal localized semantic shifts, indicating that certain regions contribute stories with distinctive cultural motifs or domain-specific content. These patterns demonstrate that, although the language is the same, regional storytelling traditions introduce measurable semantic variation.

(RQ1) How well do Large Language Models comprehend Amharic narratives?

Multiple-choice question answering (MCQA). As shown in the Figure 6 (Left), The MCQA results reveal a bimodal pattern in Amharic narrative comprehension. While Gemma models retain high accuracy on Amharic tasks (above 0.80%), all other models exhibit a sharp performance collapse to the 0.20–0.35 range, despite achieving great accuracy on English tasks. This separation suggests that cross-lingual narrative transfer is not gradual but model-dependent, likely reflecting differences in pretraining data or representational alignment rather than model size alone. Notably, English and Amharic prompting yield nearly identical performance for low-performing models, indicating a hard comprehension bottleneck where prompt language has diminishing returns. Moreover, strong English performance does not reliably predict Amharic comprehension: large multilingual models such as *Aya 23 35B* perform comparably to much smaller models on Amharic MCQA. Together, these findings indicate a structural limitation in Amharic narrative understanding rather than a failure of instruction following. Notably, *Llama 3.1 8B Instruct* exhibits the largest performance gap between English and Amharic prompting, suggesting that its narrative comprehension abilities are disproportionately stronger in English. This gap points to insufficient Amharic narrative exposure during training and motivates RQ3, where we in-

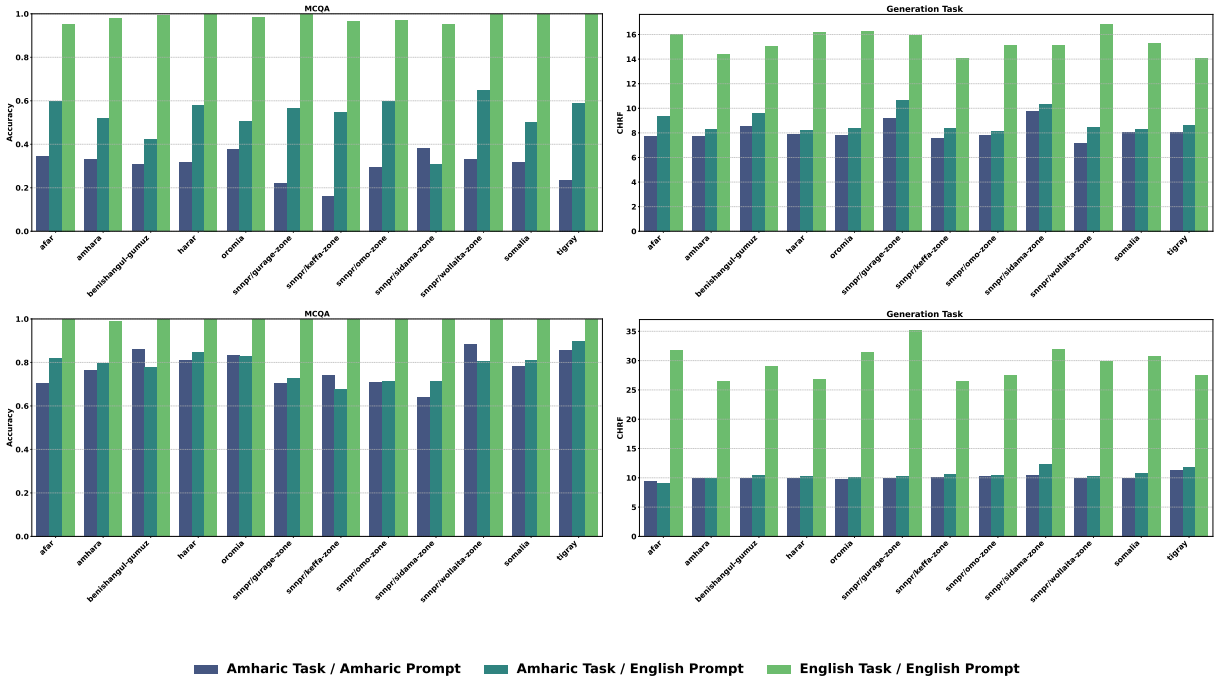


Figure 7: **Effect of SFT on *AmharicStoryQA* across regions.** Performance of *Llama 3.1 8B Instruct*(top) and its SFT variant trained on *AmharicStoryQA* (*SFT Llama 3.1 8B (All)*, bottom) on MCQA and generative evaluation.

vestigate whether increased and targeted Amharic narrative data can mitigate this disparity.

Generative comprehension. As shown in Figure 6 (right), in contrast to the sharp performance collapse observed in MCQA, generative evaluation using chrF exhibits a more gradual degradation when narratives are presented in Amharic. This pattern indicates a consistent under-performance across all Amharic results, rather than abrupt failure on specific instances. While chrF scores drop substantially from English to Amharic tasks, models retain partial generation capability, suggesting preserved lexical and syntactic knowledge despite limited narrative understanding. Providing English prompts, in most cases, for Amharic narratives yields higher chrF scores than Amharic prompts, with gains of roughly 1–3 points, indicating improved surface-level fluency. However, these improvements do not correspond to higher MCQA accuracy, highlighting a dissociation between generative fluency and true comprehension.

Across both **MCQA and generative benchmarks**, models exhibit a consistent and systematic performance degradation when operating on Amharic narratives, preserving relative model rankings and prompt-language effects, which jointly indicate a structural limitation in Amharic narrative understanding rather than benchmark-specific failure.

(RQ2) How cultural differences influence LLM performance in a single language? In Figure 3, we observe variability in sequence length across stories from different regions, highlighting region specific narrative characteristics beyond the semantic based analysis presented in Figure 2. In this research question, we aim to break down Amharic narrative understanding capabilities and examine how they are influenced by regional differences. Through manual exploration, we observe that stories often focus on region specific topics shaped by local traditions; for example, narratives from the Afar region where pastoralism is prevalent frequently involve camel herding and cattle theft.

To answer this question, Figure 5 examines two models with contrasting characteristics, *Gemma-3-27b-it* and *Llama 3.2 1B Instruct*, and analyzes their performance across regions. For English narrative understanding, we observe noticeable regional variation in most model–task settings, with differences of up to ±10 chrF points or nearly 20% accuracy. An exception is the MCQA performance of *Gemma-3-27b-it*, which remains relatively stable across regions. In contrast, Amharic narrative understanding exhibits reduced regional variance, likely due to the uniformly low baseline performance of the models. Similarly, differences arising

Model Name (Trained Split)	language	prompt_language	MCQA (accuracy)	Generation(chrf)
SFT Llama 3.1 8B (All)	English	English	**0.998**	**28.931**
SFT Llama 3.1 8B (Eng)		English	0.992	23.901
SFT Llama 3.1 8B (Amh)		English	0.988	18.023
Llama 3.1 8B Instruct		English	0.986	15.283
SFT Llama 3.1 8B (All)	Amharic	English	**0.802**	**10.410**
SFT Llama 3.1 8B (All)		Amharic	0.787	10.086
SFT Llama 3.1 8B (Amh)		English	0.758	10.056
SFT Llama 3.1 8B (Amh)		Amharic	0.748	9.758
SFT Llama 3.1 8B (Eng)		Amharic	0.540	8.259
SFT Llama 3.1 8B (Eng)		English	0.309	7.661
Llama 3.1 8B Instruct		Amharic	0.383	7.944
Llama 3.1 8B Instruct		English	0.701	8.603

Table 1: Effect of training *Llama 3.1 8B Instruct*on different *AmharicStoryQA* subsets (English-only, Amharic-only, and combined) for narrative understanding.

from prompt language choice in Amharic are attenuated in regional evaluations, suggesting that regional effects are overshadowed by broader limitations in Amharic narrative comprehension.

(RQ3) Can we improve LLM performance using the culturally grounded training dataset we introduce, demonstrating the importance of such data for addressing the limitations observed?

Using both the English and Amharic training splits of the *AmharicStoryQA* dataset, we aim to improve the narrative understanding capabilities of existing LLMs and to demonstrate that our curated dataset can mitigate the narrative comprehension limitations observed in prior evaluations. We select *Llama 3.1 8B Instruct*as our primary model due to its comparatively low and variable performance in the benchmarking results shown in Figure 6. This choice allows us to more clearly assess the impact of our training data. In contrast to models such as the Gemma family, which exhibited stronger baseline performance, *Llama 3.1 8B Instruct*provides a suitable testbed for evaluating whether targeted supervision can substantially improve Amharic narrative understanding.

As shown in Table 1, LoRA fine-tuning on the combined and English-only splits leads to clear gains in English narrative comprehension, with the largest improvements observed in generative evaluation, reaching up to +13 chrF. In contrast, MCQA performance shows limited gains, likely due to near-saturated accuracy in English settings.

For Amharic narrative understanding, we observe substantially larger benefits, with MCQA accuracy improving by nearly 40% compared to the base instruction model. Notably, the performance gap between English and Amharic prompting narrows when training on the combined dataset, indicating reduced prompt-language sensitivity. Moreover, cross-lingual training, using English stories to improve Amharic evaluation and vice versa, consistently improves both MCQA and generative scores, providing evidence of effective cross-lingual narrative transfer.

Fine-tuned models and regional improvements. Figure 7 shows that fine-tuning on *AmharicStoryQA* both improves MCQA performance and stabilizes results across regions, substantially reducing the regional disparities observed in the base model. This stabilization effect is particularly evident for Amharic narrative understanding, where fine-tuning leads to more consistent performance across diverse regional story distributions.

In contrast, regional differences are more pronounced for English narratives, even after fine-tuning. This suggests that fine-tuning primarily benefits Amharic narrative representations, while also revealing underlying regional characteristics that remain less influential for English benchmarks. As a result, fine-tuned models exhibit clearer regional variation in English tasks, indicating that regional narrative properties are more visible once the model's overall comprehension improves.

6 Conclusion and Future Work

In this work, we investigated multicultural variation within a single language and its impact on long-form Amharic story question answering. We introduced *AmharicStoryQA*, a regionally grounded benchmark designed to capture cultural and narrative diversity among Amharic speakers, and used it to systematically evaluate the narrative understanding capabilities of large language models across regions. Our results show that cultural variation within the same language meaningfully affects story comprehension, revealing limitations in current LLMs that are not captured by standard,

language-level evaluations. In particular, we find that regional narrative characteristics influence both selection-based and generative understanding, highlighting the need for culturally aware evaluation and training data. For future work, we plan to extend this single-language, multi-culture analysis to a broader multilingual and multicultural setting.

Limitations

Due to the scarcity of volunteers and the high level of manual effort required, this study does not extend the analysis to multiple languages. As a result, we are unable to determine whether the observations reported in this work generalize across languages or are specific to Amharic.

References

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, NM, USA. Association for Computational Linguistics.

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025a. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27758–27794, Suzhou, China. Association for Computational Linguistics.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025b. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27795–27829, Suzhou, China. Association for Computational Linguistics.

David Anugraha, Shou-Yi Hung, Zilu Tang, En-Shiun Annie Lee, Derry Wijaya, and Genta Winata. 2026. mr3: Multilingual rubric-agnostic reward reasoning models. In *International Conference on Learning Representations*, volume 2026, pages 149218–149272.

Israel Abebe Azime, Tadesse Destaw Belay, Dietrich Klakow, Philipp Slusallek, and Anshuman Chhabra. 2025a. Bridging the Culture Gap: A Framework for LLM-Driven Socio-Cultural Localization of Math Word Problems in Low-Resource Languages. *arXiv preprint arXiv:2508.14913*.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Thamar Solorio, and Dietrich Klakow. 2025b. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266, Albuquerque, NM, USA. Association for Computational Linguistics.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, FL, USA. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Tadesse Destaw Belay, Seid Muhie Yimam, Abinew Ayele, and Chris Biemann. 2022. Question answering classification for Amharic social media community based questions. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 137–145, Marseille, France. European Language Resources Association.

Thomas Berkane, Marie-Laure Charpignon, and Maimuna S. Majumder. 2025. LLM-based web data collection for research dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12610–12622, Suzhou, China. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into

multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced African languages? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12979–12998, Suzhou, China. Association for Computational Linguistics.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. 2025. Addressing Blind Guessing: Calibration of Selection Bias in Multiple-Choice Question Answering by Video Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3216–3246, Vienna, Austria. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: a benchmark for llms on everyday knowledge in diverse cultures and languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's Paradox of 'Low-Resource' Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Odunayo Ogundepo, Tajuddeen R. Gwadabe, Clara E. Rivera, Jonathan H. Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure F. P. Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Njoroge Kahira, Shamsuddeen Hassan Muhammad, Akintunde Oladipo, Abraham Toluwase Owodunni, and 33 others. 2023. AfriQA: Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 56 others. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *Preprint*, arXiv:2406.05967.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025. Evaluating Dialect Robustness of Language Models via Conversation Understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi, UAE. Association for Computational Linguistics.

Tilahun Abedissa Taffa, Ricardo Usbeck, and Yaregal Assabie. 2024. Low Resource Question Answering: An Amharic Benchmarking Dataset. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 124–132, Torino, Italia. ELRA and ICCL.

Sizhe Yuen, Ting Su, Ziyang Wang, Yali Du, and Adam J Sobey. 2025. Automatic Dataset Generation for Knowledge Intensive Question Answering Tasks. *arXiv preprint arXiv:2505.14212*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Dataset Distribution in Regional

Stories depicting cultural practices, daily life, and narratives from the Amhara region constitute approximately 20% of the total *AmharicStoryQA* dataset. Stories referring specifically to the Tigray and Oromia regions account for 12% and 11% of the dataset, respectively. As shown in Figure 8, the remaining regions of Ethiopia collectively represent more than 50% of the dataset.

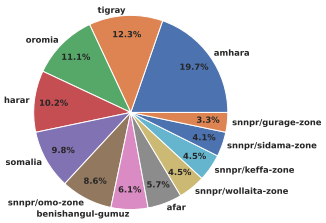


Figure 8: Dataset distribution in Region

B Ethiopian Regions

Tigray: is the northernmost regional state, bordering Eritrea to the north. It is the homeland of the Tigrayan people, who speak Tigrinya. The region is historically significant as the center of the ancient Aksumite Empire.

Oromia: is the largest regional state in Ethiopia by both population and land area, stretching across the central, western, and southern parts of the country. It is the homeland of the Oromo people, the largest ethnic group in Ethiopia, and the administrative center is Addis Ababa (Finfinne), which also serves as the national capital.

Amhara: is a regional state located in the northwestern and central highlands of Ethiopia. It is the historic homeland of the Amhara people and is known for its agricultural potential and major tourist attractions like Lalibela and Gondar. The primary language spoken is Amharic.

Somali: is the second-largest region by area, located in the eastern part of Ethiopia bordering Somalia and Djibouti. It is predominantly inhabited by the Somali people who share a common language, culture, and Islamic heritage with Somalis in neighboring countries. The region is largely arid and known for pastoralism.

Sidama: is a regional state in southern Ethiopia formed in 2020 after separating from the former SNNPR. It is the homeland of the Sidama people, who are famous for their high-quality coffee production, a major export for the country. The capital is Hawassa.

Afar: is located in the northeastern part of Ethiopia, home to the Afar people. The region is known for the Danakil Depression, one of the hottest and lowest places on Earth, and is geologically active. The inhabitants are traditionally pastoralists.

Benishangul-Gumuz: is located in western Ethiopia along the border with Sudan. It is home to diverse ethnic groups including the Berta and Gumuz.

Gambela: is situated in the western tip of Ethiopia, bordering South Sudan. It has a tropical climate and is inhabited primarily by the Nuer and Anuak peoples. The region is rich in wildlife and water resources, including the Baro River.

Harari: is the smallest regional state in Ethiopia, located in the east. It is the homeland of the Harari people and centers around the ancient walled city of Harar (Jugol), a UNESCO World Heritage site and a major center of Islamic culture in the Horn of Africa.

South West Ethiopia Peoples' (SWEP): is a regional state in southwestern Ethiopia formed in 2021. It consists of the Kaffa (the birthplace of coffee), Sheka, Bench Sheko, Dawro, and West Omo zones, representing a diverse mix of ethnic groups and languages.

Central Ethiopia: is a newly formed regional state (established in August 2023) created from the northern part of the former Southern Nations, Nationalities, and Peoples' Region (SNNPR). It is home to the Gurage, Hadiya, and Silte people, among others.

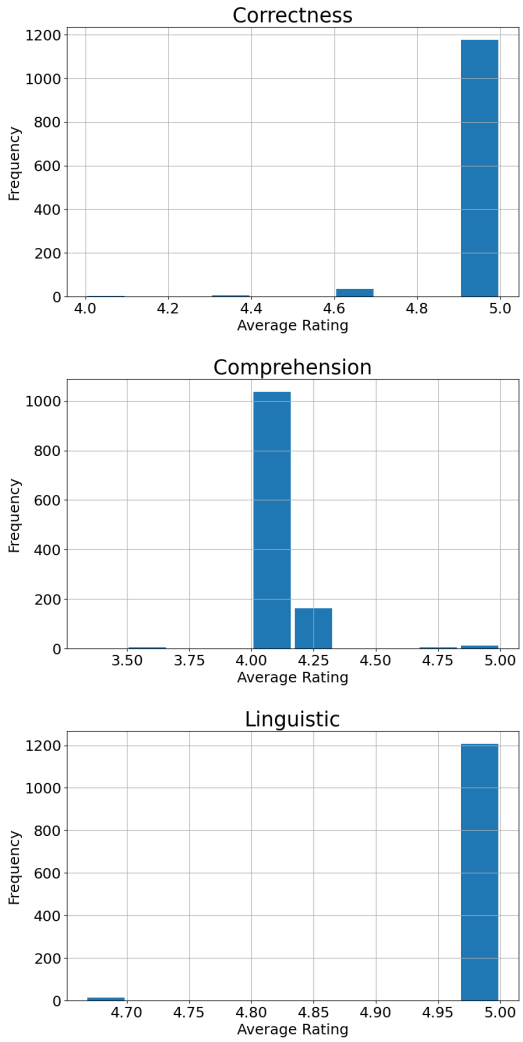
South Ethiopia: is a newly formed regional state (established in August 2023) created from the southern part of the former SNNPR. It includes zones such as Wolayta, Gamo, Gofa, and South Omo, which is famous for its diverse indigenous cultures.

Addis Ababa: is the capital city of Ethiopia and a chartered city (city administration). It serves as the headquarters for the African Union and is a melting pot of all Ethiopian ethnic groups.

Dire Dawa: is the second chartered city in Ethiopia, located in the eastern part of the country. It is a major commercial and industrial hub, historically developed around the Ethiopia-Djibouti Railway.

C Avarage Agreement Rating

Figure 9: Average rating of correctness and faithfulness, linguistic quality and clarity across three human evaluators



D Training Parameters

Fine-tuning Configuration. We fine-tune `Meta-Llama-3-8B-Instruct` using a supervised fine-tuning (SFT) stage with LoRA adapters. The training uses `lora_rank=8` with `lora_target=all` parameters while enabling `trust_remote_code`. Data preprocessing uses 16 workers and the dataloader uses 4 workers, with caching enabled and no limit on the number of samples. Training is performed with a per-device batch size of 1 and gradient accumulation of 8 steps, resulting in an effective batch size of 8. We train the model for 10 epochs using a learning rate of 1×10^{-4} with a cosine learning rate scheduler and a warmup ratio of 0.1. Mixed precision training with `bf16` is enabled. Model checkpoints are saved every 500 steps, training logs are recorded every 10 steps, and loss curves are plotted during training. All outputs are stored in the specified output directory with overwriting enabled when necessary.

E Question Generation

Question Generation

You are an expert educational content creator and an analytical text comprehension specialist. Your task is to generate challenging multiple-choice questions from the provided stories. These questions are designed to assess and diagnose language model reading comprehension.
Question Design Guidelines

- Focus on key story elements such as characters, events, causes, motivations, and outcomes.
- Ensure the question cannot be answered without understanding important parts of the story.
- Avoid vague or superficial questions.

Answer Choice Requirements
Each question must contain **four answer choices (A–D)** with the following roles:
A. Correct Answer

- The only correct option.
- Explicitly stated or directly derivable from the story.
- Requires no external knowledge.

B. Common-Sense Distractor

- Sounds plausible using real-world knowledge.
- Not supported by the story.

C. Unrelated Hallucination

- Completely unrelated to the story.
- Tests rejection of irrelevant information.

D. Factual Contradiction

- Directly contradicts a fact stated in the story.
- Must be clearly false given the text.

Input: Story: `{story}`
Output: Generate `{num_questions}` multiple-choice questions and return the result as a JSON list.

F Annotation Guidelines

Annotation Guidelines

1. Correctness and Faithfulness
Guiding Questions

- Is the question grounded in the story?
- Is the correct answer fully supported by the story?
- Do distractors follow the intended roles (plausible, opposite, irrelevant)?
- **5** Perfectly grounded in the story. Correct answer fully supported and all distractors follow their roles.
- **4** Mostly accurate with minor issues. One distractor may be weak or ambiguous.
- **3** Some factual mismatch or unclear reference. Distractors may be too similar.
- **2** Loosely based on the story. Correct answer weakly supported.
- **1** Major flaws. Distractors poorly constructed.
- **0** Completely invalid or unrelated to the story.

2. Linguistic Quality and Clarity
Guiding Questions

- Is the question grammatically correct and natural in Amharic?
- Is it fluent and easy to understand?
- **5** Excellent grammar and fluency.
- **4** Minor grammatical issues.
- **3** Understandable but awkward phrasing.
- **2** Difficult to understand.
- **1** Poor grammar and confusing phrasing.
- **0** Incomprehensible.

3. Comprehension Depth and Challenge Level
Guiding Questions

- Does the question test meaningful comprehension?
- Does it require reasoning beyond surface-level recall?
- **5** Deep reasoning or inference required.
- **4** Solid comprehension question.
- **3** Basic understanding required.
- **2** Too simple or shallow.
- **1** Trivial or irrelevant.
- **0** Completely unrelated to the task.

G Evaluation Prompt

Prompt 1 -- English:

Read the following story carefully and answer the question that follows.

Story:
"{story}"

Question:
"{question}"

Choices:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

Your answer (A, B, C, or D):

Prompt 1 -- Amharic:

የሚከተለውን ታሪክ በጥንቃቄ ያንብቡ እና ከዚያ በኋላ የሚከተለውን ጥያቄ ይመልሱ።

ታሪክ:
"{story}"

ጥያቄ:
"{question}"

አማራጮች:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

መልስዎ (A, B, C, ወይም D):

Prompt 2 -- English:

Read the story below carefully, paying attention to the details of what happens, and then answer the question that tests your understanding of it.

Story:

"{story}"

Question:
"{question}"

Choices:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

Your answer (A, B, C, or D):

Prompt 2 -- Amharic:

ከታች ያለውን ታሪክ በጥንቃቄ ያንብቡ፣ የተከሰቱትን ነገሮች ዝርዝሮች በትክክል ይጠብቁ፣ ከዚያም የገባዎትን ለመፈተሽ የሚከተለውን ጥያቄ ይመልሱ።

ታሪክ:
"{story}"

ጥያቄ:
"{question}"

አማራጮች:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

መልስዎ (A, B, C, ወይም D):

Prompt 3 -- English:

Carefully read the following passage to understand the events, emotions, and main ideas it conveys, then choose the most accurate answer to the question that follows.

Story:
"{story}"

Question:

"{question}"

Choices:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

Your answer (A, B, C, or D):

Prompt 3 -- Amharic:
የሚከተለውን ጽሑፍ በጥንቃቄ ያንብቡ እና ውስጡ የተገለጹትን ክስተቶች፣ ስሜቶች እና ዋና ሀሳቦች በትክክል ለመረዳት ይሞክሩ፣ ከዚያም የሚከተለውን ጥያቄ ለመመለስ ትክክለኛውን አማራጭ ይምረጡ።

ታሪክ:
"{story}"

ጥያቄ:
"{question}"

አማራጮች:
A) {choice1}
B) {choice2}
C) {choice3}
D) {choice4}

መልስዎ (A, B, C, ወይም D):