GAIfE: Using GenAI to Improve Literacy in Low-resourced Settings

Anonymous ACL submission

Abstract

Illiteracy is a predictor of many negative social and personal outcomes. Illiteracy rates are particularly high in countries with under-resourced languages, where few books exist that are suitable for children to learn to read from. We present GAIfE (Generative AI for Education), a toolchain and workflow developed through empirical methods, that demonstrates how existing tools can be adapted to address low literacy for an underresourced language. We used GAIfE (a play on the Bambara word for "book") to construct materials for developing children's reading competence in Bambara, the vehicular language of Mali. Our approach to the generation and post-generation editing of content skewed by the Global-North-centric bias of available LLMs, enabled us to rapidly multiply the content in Bambara available online by 10 times while maintaining high standards of attractiveness of the material to maintain high engagement, accurate representation of the Malian culture and physical and social environment and language quality. Using our materials, pilot reading programs achieved a 67% reduction in the number of children unable to read Bambara. Our approach demonstrated the power of bias-aware application of generative AI to the problem domain as well as the potential impact the application of this technology could have on reducing illiteracy and improving learning outcomes through native language education.

1 Introduction

One of the primary challenges to the development of Global South nations is illiteracy. Low literacy rates are associated with poor health, less wealth, and many other negative impacts (Cree et al., 2023). Literacy rates are particularly low among speakers of underresourced languages. These languages are often eschewed as a language of instruction, even though teaching in them is considered a more effective way to literacy to young learners (Laitin et al., 2019; Ramachandran, 2012). A lack of books in national languages constitutes one of the greatest barriers to effective instruction in mother tongue across Africa (Pflepsen et al., 2015)

Figure 1: An example of the kinds of cultural biases common in generative models that can negatively impact reading comprehension outcomes in teaching materials. The image on the left was characteristic of all images where the prompt specified a praying Muslim woman. The position of the hands is characteristic of Christian prayer, not Muslim. Our human-in-the-loop process was able to correct the image, as shown on the right. Such culturally foreign references can cause confusion among early readers and lead to poor development of reading comprehension skills. Our workflow is designed to leverage generative models to create culturally-engaging materials for learning reading, in spite of the biases present in the models.

Recent advancements in generative AI technology have shown enormous promise at performing creative tasks, particularly storytelling. However, our preliminary experiments show that, without substantial human intervention, state-of-the-art models are rather ineffective in the context of Global South storytelling for children.

Figure 1 demonstrates their limitations. In this example, we have expressly chosen an illustration where the bias, to Global North eyes, is extremely subtle, yet immensely significant. Prompts asking Stable Diffusion for images of a *praying Muslim woman* invariably showed women holding their hands together palm-to-palm, such the image on

the left. However, this manner of prayer is unfamiliar to Muslims, and in fact is characteristic of Christians. The impact of such unfamiliar imagery in early-level reading materials is that it creates confusion among young readers, because the picture does not appear match the text they are learning to read.

We developed a workflow that allowed us to use state-of-the-art generative AI tools to efficiently generate a collection of illustrated reading materials written in a language with very few books written expressly for children, written and illustrated in a manner that is culturally and age appropriate and engaging for native speakers of the language. The image on the right shows the outcome of this workflow, a woman praying in a manner that is familiar to Muslims.

We make the following contributions: (1) We introduce **GAIfE**, Generative AI for Education (inspired by "Gafe" the Bambara word for "book") a workflow for using LLMs, other AI models, lingusitic tools, and human input for constructing material designed to develop the reading and language skills of children speaking underresourced languages. The workflow produces books organized into ten reading levels, corresponding very roughly to a school grade level, thus covering a range from beginning readers through high school, written in Bambara and anchored in the physical and cultural environment of Mali.

Using GAiFE, we were able to obtain acceptable results in using AI with an intensely human-in-the-loop workflow augmented with linguistic tools to generate static educational content in Bambara using GPT4 (Achiam et al., 2023), Stable Diffusion (Rombach et al., 2022), and NLLB-200 (Team et al., 2022) models, among others. While we anticipated Ferrara (2023) and routinely encountered both egregious and subtle bias in LLMs that rendered content unacceptable according to our standards, we found that human judgement coupled with layered prompting strategies and post-generation objective-guided editing could produce acceptable results.

(2) We introduce the **GAIfE Bambara Learning Library**, a collection of 174 finished, illustrated books, 94 designed for print and 80 for electronic media such as tablets or cell phones. There are approximately 4000 pages of content, over 850 original images, and hundreds of pages of in-text questions to train reading comprehension and hundred of pages of Teaching Guides covering most of the stories. The books are all freely available in the Bloom Library[1].

Prior to our contribution to the Bloom Library, only 10 of library's contents of 19,000 books were in Bambara, of which 6 were translated Christian Bible stories for children from non-Malian sources and 4 were health information. There is a scattering of children's books in other repositories, mostly translations of non-Malian works. Printed children's books by Malian authors in Malian languages do exist in Mali, but are not commonplace and are difficult to find.

(3) We conduct a pilot study of the effectiveness of our material with teachers and students.

2 Related Work

Generative AI models such as ChatGPT have shown promise in enhancing literacy and learning outcomes, especially in high resourced environments (Ciampa et al., 2023; Dalgıç et al., 2024; Alshahrani, 2023). However, its utilization through curriculum creation to improve literacy in low-resourced settings is an emerging area with limited research since most of the world's languages are missing in the state-of-the-art generative models. Furthermore, illustrations and images generated in these contexts are Euro-centric with little to nothing to do with the realities of low-resourced settings.

Additionally, for effective AI-assisted learning, active oversight and critical evaluation of AI outputs are crucial. The complementarity of human-AI prompting strategies, such as prompting generative AI while leveraging local native speakers' unique insights and perspectives, can enhance outcomes while mitigating risks like AI biases, culture-washing, or complacency.

Han et al. (2023) propose a generative-AI-driven service, AIStory, based on focus group discussions, to help children construct visual narratives via a structured approach where children chose via a visual characters, backgrounds, props, and other story elements and the tool generates story ideas and helps the user generate images. The service also generates the images. The service does not appear to be publicly available and although the authors suggest educational impacts, that is not the express goal of the service. Nor does the service construct educational materials. Finally, the service is aimed at children themselves. This is in contrast to our approach in which adult educators

[1] https://bloomlibrary.org/RobotsMali

and computer scientists work interactively with AI systems to construct stories.

Choi et al. (2024) explore the usefulness of LLMs in the poorest schools by deploying TheTeacher.AI in Sierra Leone for teachers to utilize it in lesson planning, subject matter, and classroom management. Unlike in our work, we utilize a national, low-resource language, the language used in their study is English.

Nanduri and Bonsignore (2023) envision revitalizing endangered languages by leveraging AI-powered language learning as a catalyst for language appreciation. Their work describes the possibilities, while ours implements and evaluates in a real-world scenario.

Olson (2022) explores text, image, audio, and video modalities utilizing deep generative multimedia for children's literature. Their work is done without conducting experiments to investigate the targeted audience perceptiveness to the generated materials. In contrast to our work, we utilize iterative experiments where generated materials are field tested and improved by incorporating received feedback.

Stap and Araabi (2023) investigate how ChatGPT and other LLMs perform poorly as a translator of indigenous and extremely low-resourced languages. We observed similar results in our work in translating from high-resourced languages such as French or English to Bambara.

Challenges still remain in deploying generative AI equitable across diverse contexts and languages to ensure the quality, accuracy, and cultural appropriateness of generative AI's output (Ocker et al., 2024). Generative AI does show potential for literacy education in low-resource settings when integrated thoughtfully alongside human instruction and oversight with deliberate consideration of potential risks, biases, and ethical implications.

3 Methods

3.1 Ensuring cultural relevance from generative models

Story-writing is an important use case of generative AI, so much so that many models are fine-tuned expressly for this purpose. However, our preliminary experiments with these and other models revealed them to have significant biases against global south cultures and toward global north, colonizing cultures. Beyond the example from Figure 1, Common global north sites such as subways, snow, or blonde girls in ponytails are unseen in the Global South, making occurrences of such in text or images accompanying text confusing or meaningless, especially to children, and this can have devastating impacts on learning outcomes for improving reading comprehension.

To address such biases, we adopted a set of *fairness criteria for culturally-meaningful content for young learners*, and apply them criteria at each step of our workflow.

(1) Always present the **point of view of the target culture** in an authentic and dignified fashion, and avoid content that represents the view from of an outsider, no matter how sympathetic. In Mali, children's content is overwhelming imported from France and presents concepts that are not readily understood, impacting learning, and may be constructed with cultural assumptions that implicitly denigrate aspects of Malian life.

(2) Taken as a whole, the stories must be **inclusive** from the point of view of sex, ethnic identification, and socio-economic background. Every reader should see themselves represented and validated in at least one of our stories.

(3) Names, places, and situations must be **familiar to the target readers**.

(4) **Avoid** content that is too **topical**, timely, or centered on current political tropes.

(5) When the target audience is **children**, all care must be taken that content will not harm the children psychologically and will depict their environment in an affirming way that promotes their sense of self-worth.

(5a) Content must be emotionally resonant, entertaining and educational for children in the target culture.

(5b) Warmth, love, positivity, and reinforcement of self-esteem should under-gird all content. Challenging content may be presented, but always in a way that promotes personal development and compassion.

(5c) Avoid metaphors and imagined and abstract language that can only be understood with the assumption that the child has a background in another culture, particularly in the culture of the former colonizing power.

(5d) Graded material at different levels should reflect the intellectual development of children at different ages and backgrounds in the target culture. Themes and images must always be age-appropriate.

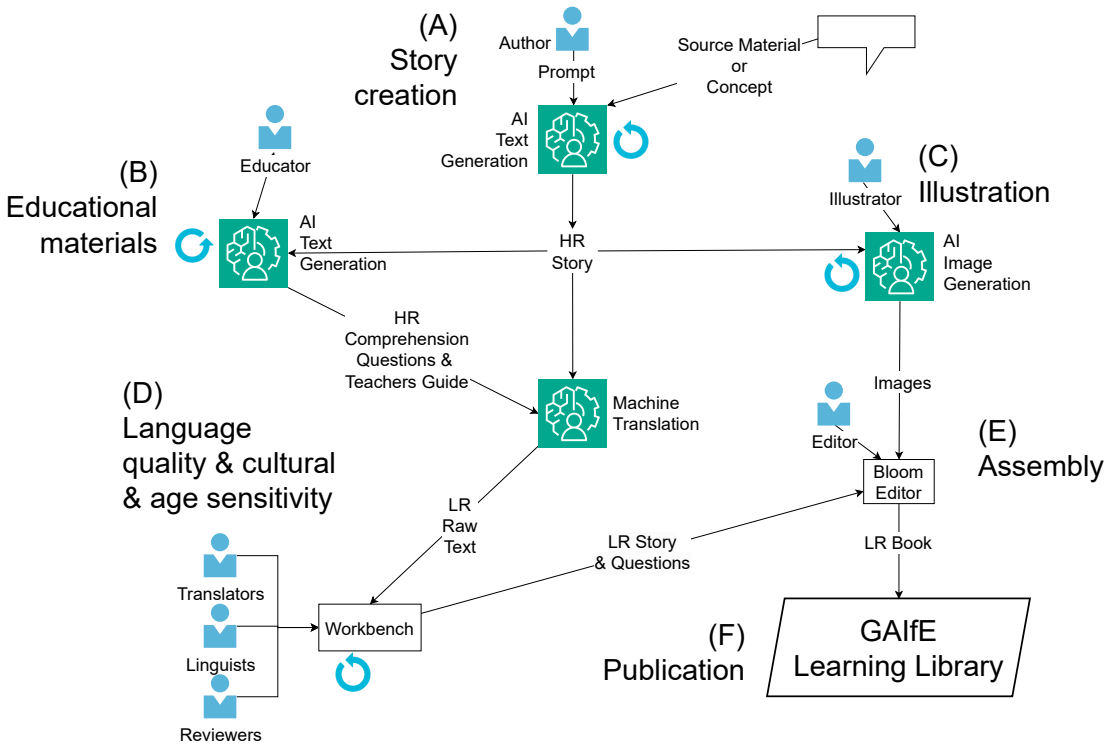

Figure 2: The GAIfe workflow leverages the creative power of high-resource (HR) languages to generate stories that are (A, B, C) iteratively refined to make them culturally and age appropriate. They are then converted to a low-resource (LR) language using state-of-the-art LLM-based machine translation. After a final round (D) of manual editing, they are assembled (E) and published (F).

3.2 GAIfE Workflow

We start (Figure 2 (A)) with a *writer* in the target language (such as Bambara) using ChatGPT (OpenAI, 2023) to convert a story idea into a story text in either English or French. This typically involves a great deal of manual prompt-tuning to get ChatGPT to produce stories that respect our cultural and language guidelines. Once a story has been generated we also generate comprehension and vocabulary questions, which are included at the end of the story, and lesson materials for teachers with activities to encourage skill development and exploration of the themes and values encountered in the story. This step typically requires less human intervention as the source material used for generation has already been shaped into a story rooted in the Malian environment.

Once the story is fixed, an illustrator decides on the set of images, trying an initial prompt for each target image, using (B) a text-to-image generator. A single prompt rarely produces the desired image, prompts typically are continuously tweaked in an attempt to find the right prompt elements that will converge toward a good image. Often a base prompt will be abandoned and another approach toward describing the image will be tried. If an image begins to approach the desired outcome but needs further modification, the illustrator may switch to image-to-image generation where prior images form the prompt for generating a new image. It is not unusual that a thousand or more images would be generated before obtaining an acceptable result, where each generated image must be reviewed by the illustrator. The tools that we used were not able to respond in consistent and predictable ways to our prompts, forcing us to rely on random variation over a large number of outputs to produce suitable images in a process that converged toward the desired image. We use Stable Diffusion in the Playground AI (pla) interface, MidJourney, DallE2 and DallE3, sometimes in combination, along with other image editing tools. We found that each model has distinct strengths and weaknesses. Stable Diffusion was the model we employed the most due to its speed and the lack of superfluous elements that other models typically added, giving images, to our eyes, a less realistic appearance and a characteristic “AI-generated“

look. Once all images are deemed appropriate they will incorporated into a storybook in (E).

Simultaneously, the text is translated into the target language (C) using Google Translate or Glosbe. Next (D), the Bambara translation, along with the source, are loaded into Annotator, a linguistic-analysis tool chain and text-annotation tool purpose-built for this workflow (see Figure 4 for a screenshot of this tool). Humans review the translation for appropriateness. The translation is corrected, adjusted, and reviewed until the Bambara language is fluid, colloquial, and appropriate to the target age of the reader.

In step (E), we use the Bloom Editor (blo) to collate the story with its corresponding illustrations. Finally (F), we produce the book and send it into the field for testing. The finished book is tested and feedback are collected to improve the books if need arises. In the field we evaluate the appropriateness of the content with children, teachers and parents as well as effectiveness in terms of learning outcomes and teacher adoption.

3.3 Prompt Analysis

We were able to generate a large quantity of text, images and pedagogical material using generative AI, but, we were not able to generate the content meeting our standards without refining content through successive prompts shaping the material, replacing biased representations with accurate representation of the target culture and environment and preparing the language for efficient automatic translation that would minimize rewriting. We analyzed the prompt history for generation of story, comprehension question, and Teacher Guide text for 10 books, classifying the objective of each type of prompt. We have made the full analysis and the prompts themselves with the initial generated story text and its final version, before translation, available in a repository. In this section, we limit the discussion to the story generation prompts.

We classified prompts according to 7 different target objectives as follows: (i) Ideation, explore a knowledge domain to help in forming ideas for stories; (ii) Generation Instruction from Seed, create book content. Seed may include main idea of story, setting, language level, writing style; (iii) Story Shaping, add, replace or delete narrative elements of the story, either to better fit content objectives or for aesthetic reasons; (iv) Cultural Adaptation, changes to content specifically to make the story fit a Malian context, typically originating from model bias and in spite of directives in the seed; (v) Language Level Adjustment, changes to vocabulary to better fit the story to the target age level, typically necessary because of model bias and in spite of directives in the seed; (vi) Language Idiom Adjustment, changes to expressions used in the text to avoid idioms that will be difficult to translate into the target language or will not be understood if translated literally, typically necessary because of model bias and in spite of directives in the seed; (vii) Stylistic Improvement, changes to minor elements in the story not affecting the overall narrative to improve the story text, either for purposes of accuracy in the Malian context, simplification, or aesthetics.

The 10 books have a total of 256 pages, including front and back matter, images, and a section of comprehension questions. The reading level of the books was from B to K on an A to K scale, with an average reading level of F, corresponding, to a wide approximation, to the target level of childen in Mali between 8 and 10 years of age. A single prompt might contain several instructions acting on different aspects of the text and with different purposes, so in our analysis we identified and counted all the discrete instructions contained in a prompt field. The average number of instructions per book was 28.9, or 1.1 instruction per page. For story content, there were 16.6 instructions, on average, per book. 30% of the story instructions were motivated by cultural adaptation, arising from the inability of the LLM to represent a Malian environment despite explicit directives in the generation from seed instructions. 6.6% of story instructions arose from the need to replace idiomatic language based on cultural knowledge unfamiliar to Malian children. 22.3% of story instructions were ascribed to the LLM failing to use vocabulary and concepts that were at target age-level. We believe that some of this should be ascribed to biased training data. As a rough estimation, about half of the prompts were related in some way to bias, the other half being more of the nature of refining the storytelling. Our authors felt the process was extremely efficient, a subjective view bolstered by the data showing less than 30 instructions, on average, being needed to craft an excellent story specifically tuned to the worldview of Malian children.

A summary of the breakdown of prompt instructions for story generation for the 10 books is shown below.

Instruction Type - Story Content	Total Instructions Type	Total Story Instructions (%)	Avg. # of Type/Book	Avg. # of Type/Page
Ideation	4	2.4	0.4	0.02
Generation Instruction from Seed	13	7.8	1.3	0.05
Story Shaping	21	12.7	2.1	0.08
Cultural Adaptation	50	30,1	5	0.20
Language Level Adjustment	37	22.3	3.7	0.14
Language Idiom Adjustment	11	6.6	1.1	0.04
Stylistic Improvement	30	18.1	3	0.12

Table 1: The story content prompts types and statistics

4 GAIfE Bambara Learning Library

We used GAIfE to create the GAIfE Bambara Learning Library, a collection of 174 finished, illustrated books, 94 designed for print and 80 for electronic media such as tablets or cell phones. 16 of the books are in Malian national languages other than Bambara, while the rest are in Bambara. There are approximately 4000 pages of content, over 850 original images, and hundreds of pages of in-text questions to train reading comprehension and hundred of pages of Teaching Guides covering most of the stories. The books are all freely available in the Bloom Library[2].

Through our workflow, we were able to construct this library in approximately 6 months. Malian authors and illustrators worked with technical experts to create the books. Apart from a few experiments where we allowed generative AI to propose and develop a story with minimal human intervention, all of the stories begin as a product of the imagination of a creative Malian with profound knowledge of the culture and environment of Mali. Our experience proved the necessity of this approach, as nothing that we created in the low-human intervention experiments produced anything that passed subsequent evaluation of representation and relevance specific to Mali. The seed material for the stories came from a variety of sources including traditional Malian tales, Malian religions, contemporary issues of interest to children, Malian family life, and tales from world literature adapted for the Malian environment.

The Malian author used ChatGPT (OpenAI, 2023) as an accelerator and aid to the creative process, as a writing tool, and as a research assistant. We confirmed that ChatGPT is unable to generate intelligible content directly in Bambara and therefore had the authors work in either French or English. Despite our inability to work directly in Bambara, we found the use of ChatGPT to significantly increase the speed with which content was generated and its quality.

For example, the book shown in Figure 3, "Mani ni Bama"[3], was created as an abridgement of a long story in French by the renowned Malian storyteller, Awa Bakɔrɔba Dembélé. It was decided to produce a version that would be accessible to younger readers. The "reduction" process involved having ChatGPT identify the essential elements of the story and limiting the vocabulary and simplifying the construction and grammar of the sentences. A very good result was obtained in a few minutes - in the same time the author, trying to produce a hand-crafted reduction, had almost completed the first sentence of the derivation.

Many stories are based on external sources, such as "Uncle Tom's Cabin"[4] and Chinese classic "Dream of the Red Chamber""[5]. ChatGPT was used by the author to help narrow down relevant elements from the source of inspiration and to smoothly incorporate them into the storyline at the appropriate level of language and narrative complexity and to integrate the story into a Malian context.

ChatGPT also proved extremely useful in accurately incorporating scientific and technical elements in many stories including ocean raft building in "Taama Laban" (Taa), and metallurgy in "Ntanan Dilanna"[6]. Creation of a story aided by ChatGPT typically took about half a day. The authors participating in the project estimated that without ChatGPT the same production would have taken weeks

[2] https://bloomlibrary.org/RobotsMali

[3] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluI/book/JZpFSC0HnB

[4] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluH/book/0UBurh0ytj

[5] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluG/book/qTmdvelbYG

[6] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluK/book/4miNF2NgDQ

Figure 3: An example of book cover for the story of "Many ni Bama".

or, in all likelihood, would not have been undertaken at all given the difficulty of assemblying and analyzing the various sources used.

The remaining phases of the workflow required similar adaptions for working with the resources available for supporting Bambara and are detailed in the Appendix.

5 Evaluation

To evaluate the effectiveness of the education materials in enabling children to learn to read Bambara and to address illiteracy, we set up six reading programs, 3 in urban and 2 in rural community centers and 1 in a school. Specifically, we wanted to assess if the texts were age-appropriate and comprehensible for the development level of the child, if the stories delighted the children, and if instructors were able to use the materials for teaching language skills in informal and formal learning environments.

We collected $m = 475$ responses from $n = 300$ unique participants (some of whom participated in multiple sections). The ages of respondents ranged from 4 to 15, with the median age being 9. All participants were native Bambara speakers. With a handful of exceptions, the children at the urban community centers and the school had been taught to read in French, and none knew Bambara letters or how to decode words written in Bambara. In one of the rural community centers, the children had been taught to read letters of the Bambara alphabet and how to decode words written in Bambara, but not to read full sentences (due to lack of reading materials), and they had not learned French. In the other rural community center, in Safo, all participants were completely illiterate in any language, the majority having attended little or no school. In the rural centers, almost all parents of the participants were illiterate.

The reading programs all consisted of three half-day sessions, or approximately 12 hours of activities, with the exception of Safo where the program extended over 8 half-day sessions for a total of 32 hours of activities.

We start by asking some demographic questions, of both teachers and students, such as: age, sex, languages spoken, languages written.

The sessions' main activity consisted of a group activity where children read collectively and individually GAIfE books, mainly in printed form but also, at some sites, online and answered the comprehension questions that accompanied each book. There was also guided discussion about the stories and considerable improvisation on the part of the teachers. All sessions begin with an alphabet primer[7].

The ability of each student to read Bambara was tested individually with a simple exercise asking them to read Bambara letters and to recognize simple words that any Bambara-speaking child would know orally such as chicken, fish, or dog. The progress of each student was noted by the instructor during the sessions based on the student's individual reading and responses to comprehension questions. At the conclusion of the reading program, students were evaluated on reading ability by being presented with the highest level text they were capable of reading and observing whether they were able to read the text fluidly and respond to comprehension questions requiring both ability to read the question and to have understood details in the story they just read.

All children were evaluated to be unable to read in Bambara at the beginning of the reading sessions. Children that had learned to read in French demonstrated a latent capacity to recognize words in Bambara that could be approximated using French pronunciation but were unable to read many words and full sentences. Once they were able to read complete books with full sentences at any pace

[7] https://bloomlibrary.org/player/LDTS2VfmHH

we noted them as being able to read. All school-age children that had not learned to read in any language were classified as illiterate. The same criteria as used with French-capable children was employed in designating them as no longer being illiterate. At the end of either the 12-hour or 32-hour programs we measured the following improvement:

- Urban Community Reading Programs reduction of inability to read in Bambara by 53 %
- Rural Community Reading Programs reduction of illiteracy rate by 79 %
- School-based Reading Program reduction of inability to read in Bambara by 70 %

Most children learned to read in a very short time using the books produced by the GAIFE project. There are numerous variables in these experiments and we did not set up the experiments to allow us to perform multivariate regression analysis on the results. The possible elements at play include: the books themselves, including the selection of themes, the resonance of the books for Malian children, the illustrations, and the interactive elements, the novelty of books in Bambara as the most of the children had never seen a book in their mother tongue and had little experience of children's books in any language, the pedagogical approach which aimed to make the reading sessions fun, in marked contrast to the focus on discipline in a typical Malian school setting, and that the children were given printed books to take home, conferring a concrete value on participation in the reading session, and the pre-literacy skills of the children, which were substantial given their average age of 9.

5.1 Parent and Teacher Perspectives

Teachers and parents were surveyed to assess their attitudes toward our materials and the idea, in general, of promoting reading in the children's native tongue. See Tables 3 and 4 in the appendix for more details about the perspectives of parents and teachers in urban versus in rural areas. In urban environments, the eagerness and receptivity of the children stood in marked contrast to the skepticism of many parents and teachers. The majority of urban parents being French-speaking, the concern of parents may be explained by the fact that education in Mali is currently entirely in French and that many jobs in the formal sector require French-language competence. These results stand in marked contrast with the responses from parents in rural settings, where fully 100% of interviewed parents were eager for the children to learn to read Bambara. One parents succinctly expressed the view of this cohort, "Why should our children learn French, it isn't our language. We have our own language."

6 Conclusion and Future Work

We were able to use generative AI to create a significant quantity of children's stories and accompanying pedagogical material in Bambara, an under-resourced language, in a short period of time. While ChatGPT is unable to directly generate intelligible Bambara text during our investigation, generating text in English or French followed by translation using Google Translate and Glosbe proved to be an efficient process, though one that requires expert-level human intervention with cultural appropriateness knowledge and context. Human intervention was also required to remove references that are culturally inappropriate and to add elements that reflect the target culture. Children proved to be receptive, showing strong motivation to read, for the first time, in their native language. Almost all, in a short period of time, did demonstrate that they possessed a high degree of latent ability to read in Bambara. Additionally, they were able to demonstrate excellent comprehension of the stories. In contrast, parents and teachers in urban environments expressed skepticism about the value of reading in Bambara.

More assessment is crucial for the next phases of this seminal work for predominantly oral languages (POLs). These techniques need to be deployed to a wider audience of content creators, in Mali and elsewhere in the world. This has already begun, our team having trained national language educational specialists in the Malian government and having shared our methods and results with the community around the Bloom Library and the mEducation Alliance. Addressing the eurocentrism in existing LLMs and other generative models is necessary objective to advance the state-of-the-art beyond what could be accomplished in project GAIFE. Improvement in the accuracy of translations for low-resource languages and the inclusion of many more of Mali's and the world's languages is also a primary concern for the future.

7 Limitations

Digital technologies and AI hold promise as tools that could contribute to solutions reducing illiteracy, but the possibility of their use would appear extremely limited where the languages that people speak are underresourced, as is the case in Mali where most people understand only Bambara or other languages of the region. A majority of the population being illiterate ensures that their language will remain low-resourced as a language becomes highly resourced by virtue of having a strong digital presence. If a way is not found to break this vicious cycle, the people that speak low-resourced languages appear to be doomed to remain excluded from the arc of progress technology development has brought to other parts of the world.

The high illiteracy rate in Mali is harmful in many ways. By providing materials for teaching children how to read Bambara, our work has the potential to yield enormous benefits by increasing literacy throughout Mali and thus reducing these harms. Additionally, since Bambara and its dialects are spoken in the neighboring countries, this has also the potential to be utilized by them. Furthermore, the content we generated could be leveraged to bridge the lack of readily digitized data for low-resourced languages.

There are numerous ethical challenges posed by the use of generative models for composing creative content. For one, the models used here—as well as every other foundational model that we are aware of—draw from the intellectual property of creators who never consented to having their material used to train the models, and so the use of generative models arguably constitutes an unauthorized use of their work, one that is generally impossible to attribute or compensate. Beyond that, AI models have their own biases. In spite of the fact that we provide the prompts and conduct extensive editing on the back end, the models provide the bulk of the content and this effectively frames the content in ways that may be hard to perceive or correct. Regarding generated images, they have the potential to unduly impress or even traumatize children when they contain defects that are not representative of how humans or animals look like in real life.

Moreover, our group is very small and cannot possibly represent the entire population of Mali. For instance, the writers of the material were all men. Although we believe ourselves to be good faith actors, we undoubtedly hold unconscious biases that impact our thinking and beliefs, and we lack the breadth of lived experiences that a larger population of authors would have. To have such a small group of people be responsible for such a relatively large proportion of all extant children's literature in Mali potentially would give us an inordinate amount of power to influence the thinking of the youth of Mali. We hope that by publishing our methods, other, diverse groups of authors will be inspired to amplify their voices in the same way.

To expand the discussion on the ethical implications of using AI-generated content, particularly focusing on intellectual property rights and the consent of original content creators (which is currently in the last two paragraphs of Limitations, ss 6.1), we tend to make ethical considerations using the three principles of the Belmont Report as a framing device, and so we organize our comments around them, mentioning ACL ethics principles where appropriate. From the Belmont principle of Beneficence, using large language models to create content risks having the intellectual property of others appearing (sometimes completely intact) in our work. Moreover, by using LLMs we are depriving content creators of opportunities to be paid for the work the LLMs are doing (though no one has yet paid anyone to create story-based teaching materials for reading Bambara), and contravening ACL principle 1.5. We mitigated these risks (thus avoiding harm per ACL ethics principle 1.2) by doing a substantial amount of manual editing of the content and using prompts for the LLMs and image generators that were very specific to Mali (and very little content on the web derives from Mali), essentially adding a large amount of content ourselves. Our work provides no benefits whatsoever to those content creators whose data is used to train the models. Moreover, since the LLMs we used are derived in part from the intellectual property of people who may not have consented to have their work used, this contravenes the Belmont principle of Respect for Persons. However, we believe this work has the potential to help millions of people read better in their first language, and literacy has enormous social and personal benefits, so from the perspective of the Belmont principle of Justice, and ACL principle 3.1, the benefits to society outweigh the risks by a great measure, in our view.

References

Bamadaba - Dictionnaire Bambara — cormand.huma-num.fr. http://cormand.huma-num.fr/Bamadaba/lexicon/index.htm. [Accessed 15-10-2024].

BloomLibrary — bloomlibrary.com. https://www.bloomlibrary.com/. [Accessed 15-10-2024].

BloomLibrary — bloomlibrary.org. https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-KingSidikiStories-Print/book/MIYqW8FJmc. [Accessed 15-10-2024].

Free AI Design Tool: Logos, T-Shirts, Social Media - Playground — playgroundai.com. https://playgroundai.com/. [Accessed 15-10-2024].

LLACAN | Institut National des Langues et Civilisations Orientales — inalco.fr. https://www.inalco.fr/en/llacan. [Accessed 15-10-2024].

2023. Decolonizing nlp for "low-resource languages": Applying abebe birhane's relational ethics. *GRACE: Global Review of AI Community Ethics*, 1(1).

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *Preprint*, arXiv:2203.08351.

Ninoh Agostinho Da Silva, Tunde Oluwaseyi Ajayi, Alexander Antonov, Panga Azazia Kamate, Moussa Coulibaly, Mason Del Rio, Yacouba Diarra, Sebastian Diarra, Chris Emezue, Joel Hamilcaro, Christopher M. Homan, Alexander Most, Joseph Mwatukange, Peter Ohue, Michael Pham, Abdoulaye Sako, Sokhar Samb, Yaya Sy, Tharindu Cyril Weerasooriya, Yacine Zahidi, and Sarah Luger. 2023. Findings from the Bambara - French machine translation competition (BFMT 2023). In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 110–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

A Alshahrani. 2023. The impact of chatgpt on blended learning: Current trends and future research directions. *International Journal of Data and Network Science*, 7(4):2029–2040.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

A Bapna et al. 2022. Building machine translation systems for the next thousand languages. may 16. *arXiv preprint arXiv:2205.03983*.

Aras Bozkurt. 2023. Unleashing the potential of generative ai, conversational agents and chatbots in educational praxis: A systematic review and bibliometric analysis of genai in education. *Open praxis*, 15(4):261–270.

Erika Mariana Carvajal Soriano, Antonia Abigail Cortés Pérez, María José Carvajal Carrera, Beatriz Adriana Sabino Moxo, and José Alberto Márquez Domínguez. 2024. Design of a software prototype for listening comprehension of the ixcatecan language. In *Proceedings of the XI Latin American Conference on Human Computer Interaction*, CLIHC '23, New York, NY, USA. Association for Computing Machinery.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Xiang'Anthony' Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, Chien-Sheng Wu, et al. 2023. Next steps for human-centered generative ai: A technical perspective. *arXiv preprint arXiv:2306.15774*.

Jun Ho Choi, Oliver Garrod, Paul Atherton, Andrew Joyce-Gibbons, Miriam Mason-Sesay, and Daniel Björkegren. 2024. Are llms useful in the poorest schools? theteacher.ai in sierra leone. *Preprint*, arXiv:2310.02982.

Katia Ciampa, Zora M Wolfe, and Briana Bronstein. 2023. Chatgpt in education: Transforming digital literacy practices. *Journal of Adolescent & Adult Literacy*, 67(3):186–195.

Pedro Conceição. 2019. *Human Development Report 2019*. UNESCO.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Anthony Cree, Andrew Kay, and June Steward. 2023. The economic and social cost of illiteracy: A snapshot of illiteracy in a global context. *World Literacy Foundation*, pages 1–18.

Ali Dalgıç, Emre Yaşar, and Mahmut Demir. 2024. Chatgpt and learning outcomes in tourism education: The role of digital literacy and individualized learning. *Journal of Hospitality, Leisure, Sport & Tourism Education*, 34:100481.

Darren A DeWalt, Nancy D Berkman, Stacey Sheridan, Kathleen N Lohr, and Michael P Pignone. 2004. Literacy and health outcomes: a systematic review of the literature. *Journal of general internal medicine*, 19:1228–1239.

Abdramane Diarra. 2020. *Le Curriculum Bilingue dans l'enseignement fondamental au Mali: Etat des lieux de sa mise en œuvre*. Ph.D. thesis, Université Grenoble Alpes [2020-....].

Utkarsh Dwivedi, Salma Elsayed-Ali, Elizabeth Bonsignore, and Hernisa Kacorri. 2024. Exploring ai problem formulation with children via teachable machines. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Florina Erbeli, Elsje van Bergen, and Sara A Hart. 2020. Unraveling the relation between reading comprehension and print exposure. *Child Development*, 91(5):1548–1562.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

D Ganesh, M Sunil Kumar, P Venkateswarlu Reddy, S Kavitha, and D Sudarsana Murthy. 2022. Implementation of ai pop bots and its allied applications for designing efficient curriculum in early childhood education. *International Journal of Early Childhood Special Education*, 14(3).

David Giguere and Erika Hoff. 2023. Bilingual children's vocabulary skills at 5 years predict reading comprehension development within, not across, languages. *International Journal of Bilingual Education and Bilingualism*, pages 1–13.

Henner Gimpel, Kristina Hall, Stefan Decker, Torsten Eymann, Luis Lämmermann, Alexander Mädche, Maximilian Röglinger, Caroline Ruiner, Manfred Schoch, Mareike Schoop, Nils Urbach, and Steffen Vandrik. 2023. Unlocking the power of generative ai models and systems such as gpt-4 and chatgpt for higher education: A guide for students and lecturers. Hohenheim Discussion Papers in Business, Economics and Social Sciences 02-2023, Stuttgart. Urn:nbn:de:bsz:100-opus-21463.

Yacouba Goita and Abdramane Kone. 2023. The primary education evaluation systems in mali and china must be compared in order to determine the best strategy to employ. *Education Journal*, 12(3):99–102.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Muhammad Hakiki, Radinal Fadli, Agariadne Dwinggo Samala, Ade Fricticarani, Popi Dayurni, Kurniati Rahmadani, Ayu Dewi Astiti, and Arisman Sabir. 2023. Exploring the impact of using chat-gpt on student learning outcomes in technology learning: The comprehensive experiment. *Advances in Mobile Learning Educational Research*, 3(2):859–872.

Ariel Han, Zhenyao Cai, Seungmin Jeong, and Sun Myung Choi. 2023. Aistory: design implication of using generative arts ai for visual storytelling.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

David D Laitin, Rajesh Ramachandran, and Stephen L Walter. 2019. The legacy of colonial language policies and their impact on student learning: Evidence from an experimental program in cameroon. *Economic Development and Cultural Change*, 68(1):239–272.

Bernardo Leite, Tomás Freitas Osório, and Henrique Lopes Cardoso. 2024. Fairytaleqa translated: Enabling educational question and answer generation in less-resourced languages. *Preprint*, arXiv:2406.04233.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. *arXiv preprint arXiv:2210.14712*.

Michael Leventhal, Allahsera Tapo, Sarah Luger, Marcos Zampieri, and Christopher M Homan. 2020. Assessing human translations from french to bambara for machine learning: a pilot study. *arXiv preprint arXiv:2004.00068*.

Kirill Maslinsky and Valentin Vydrin. 2019. Daba software for written corpora of underresourced languages. In *Journées scientifiques" Linguistique informatique, formelle et de terrain"/Scientific meeting of the "Computational, formal and field linguistics" research group*.

Margaret J McLaughlin, Katherine E Speirs, and Edmond D Shenassa. 2014. Reading disability and adult attained education and income: Evidence from a 30-year longitudinal study of a population-based sample. *Journal of Learning Disabilities*, 47(4):374–386.

Suzanne E Mol and Adriana G Bus. 2011. To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin*, 137(2):267.

Dinesh Kumar Nanduri and Elizabeth M Bonsignore. 2023. Revitalizing endangered languages: Ai-powered language learning as a catalyst for language appreciation. *arXiv preprint arXiv:2304.09394*.

Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, and Maggie Shen Qiao. 2021. Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041.

Davy Tsz Kit Ng, Wanying Luo, Helen Man Yi Chan, and Samuel Kai Wah Chu. 2022. Using digital story writing as a pedagogy to develop ai literacy among primary students. *Computers and Education: Artificial Intelligence*, 3:100054.

Rosalie J. Ocker, David Fusco, Edward Glantz, Steven Haynes, Lisa Lenze, Dan Welch, and Qingyun Wu. 2024. Report from the generative artificial intelligence across-the-curriculum task force.

Chinasa T Okolo. 2023. The promise and perils of generative ai: Case studies in an african context.

Matthew L Olson. 2022. Deep generative multimedia children's literature. *arXiv preprint arXiv:2209.13129*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alison Pflepsen, C Benson, C Chabbott, and A van Ginkel. 2015. Planning for language use in education: Best practices and practical steps to improve learning outcomes. *USAID Bureau for Africa*. https://www.researchgate.net/publication/281823056_Planning_for_Language_Use_in_Education_Best_Practices_and_Practical_Steps_to_Improve_Learning_Outcomes.AccessedMarch, 8:2016.

Rajesh Ramachandran. 2012. Language use in education and primary schooling attainment: Evidence from a natural experiment in ethiopia. *IEB Working Paper 2012/34*.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

RobotsMali. Bloom library. https://bloomlibrary.org/RobotsMali.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Keith E Stanovich. 2009. Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education*, 189(1-2):23–55.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. 2020. Neural machine translation for extremely low-resource african languages: A case study on bambara. *arXiv preprint arXiv:2011.05284*.

Nllb Team, Marta Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Ram Kaushik, Dirk Sadagopan, Shannon Rowe, Chau Spruit, Pierre Tran, Andrews, Necip Fazil, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and Meta Ai. 2022. No language left behind: Scaling human-centered machine translation.

Aminata Thiam. 2023. L'introduction des langues nationales dans le système éducatif du mali : enjeux, défis, et perspectives.

National Reading Panel (US), National Institute of Child Health, and Human Development (US). 2000. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National ....

Elsje van Bergen, Margaret J Snowling, Eveline L de Zeeuw, Catharina EM van Beijsterveldt, Conor V Dolan, and Dorret I Boomsma. 2018. Why do children read more? the influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, 59(11):1205–1214.

Valentin Vydrin. 2013. Bamana reference corpus (brc). *Procedia - Social and Behavioral Sciences*, 95:75–80. Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC2013).

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Weipeng Yang. 2022. Artificial intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence*, 3:100061.

A Appendix

A.1 Text Generation

Our process begins with a Malian author with an idea for a children's story. Apart from a few experiments where we allowed generative AI to propose and develop a story with minimal human intervention, all of the stories begin as a product of the imagination of a creative Malian with profound knowledge of the culture and environment of Mali. Our experience proved the necessity of this approach, as nothing that we created in the low-human intervention experiments produced anything that passed subsequent evaluation of representation and relevance specific to Mali. The seed material for the stories came from a variety of sources including traditional Malian tales, Malian religions, contemporary issues of interest to children, Malian family life, and tales from world literature adapted for the Malian environment.

The Malian author used ChatGPT (OpenAI, 2023) as an accelerator and aid to the creative process, as a writing tool, and as a research assistant. We confirmed that ChatGPT is unable to generate intelligible content directly in Bambara and therefore had the authors work in either French or English. Despite our inability to work directly in Bambara, we found the use of ChatGPT to significantly increase the speed with which content was generated and its quality. For example, the book shown in Figure 3, "Mani ni Bama"[8], was created as an abridgement of a long story in French by the renowned Malian storyteller, Awa Bakɔrɔba Dembélé. It was decided to produce a version that would be accessible to younger readers. The "reduction" process involved having ChatGPT identify the essential elements of the story and limiting the vocabulary and simplifying the construction and grammar of the sentences. A very good result was obtained in a few minutes - in the same time the author, trying to produce a hand-crafted reduction, had almost completed the first sentence of the derivation. Many stories are based on external sources. ChatGPT was used by the author to help narrow down relevant elements from the source of inspiration and to smoothly incorporate them into the storyline at the appropriate level of language and narrative complexity and to integrate the story into a Malian context. Some examples of books produced by this process include "Bɛnkɛ Tɔm Ka So"[9] derived from the American classic "Uncle Tom's Cabin" and "Gerenad-Feerew"[10] based on a tale in the Chinese classic "Dream of the Red Chamber". ChatGPT proved extremely useful in accurately incorporating scientific and technical elements in many stories including ocean raft building in "Taama Laban" (Taa) and metallurgy in "Ntanan Dilanna"[11]. Creation of a story aided by ChatGPT typically took about half a day. The authors participating in the project estimated that without ChatGPT the same production would have taken weeks or, in all likelihood, would not have been undertaken at all given the difficulty of assemblying and analyzing the various sources used.

In our work with ChatGPT we never saw evidence of the LLM having a knowledge base of the Malian culture or physical and social environment. More significant, ChatGPT would often generate story elements that, in the judgement of authors, was a false, sometimes fantastical, and, often, prejudicial view of Mali. For example, the early reader book "Anw Bɛɛ Baarakɛ¡' describes in simple language 30 different common types of work in Mali. The author asked ChatGPT to produce a list of different jobs and a short description of each in simple language accessible to a beginning reader. The results produced many types of jobs that are common in developed countries but uncommon in Mali, described jobs that are common in a way that differed so much from Malian reality that the description would make no sense to children, and did not list jobs that are common in Mali but uncommon in developed countries. Carpenter, metalworker, maid, doorman and sanitary works are example of jobs

[8] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluI/book/JZpFSC0HnB

[9] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluH/book/0UBurh0ytj

[10] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluG/book/qTmdvelbYG

[11] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluK/book/4miNF2NgDQ

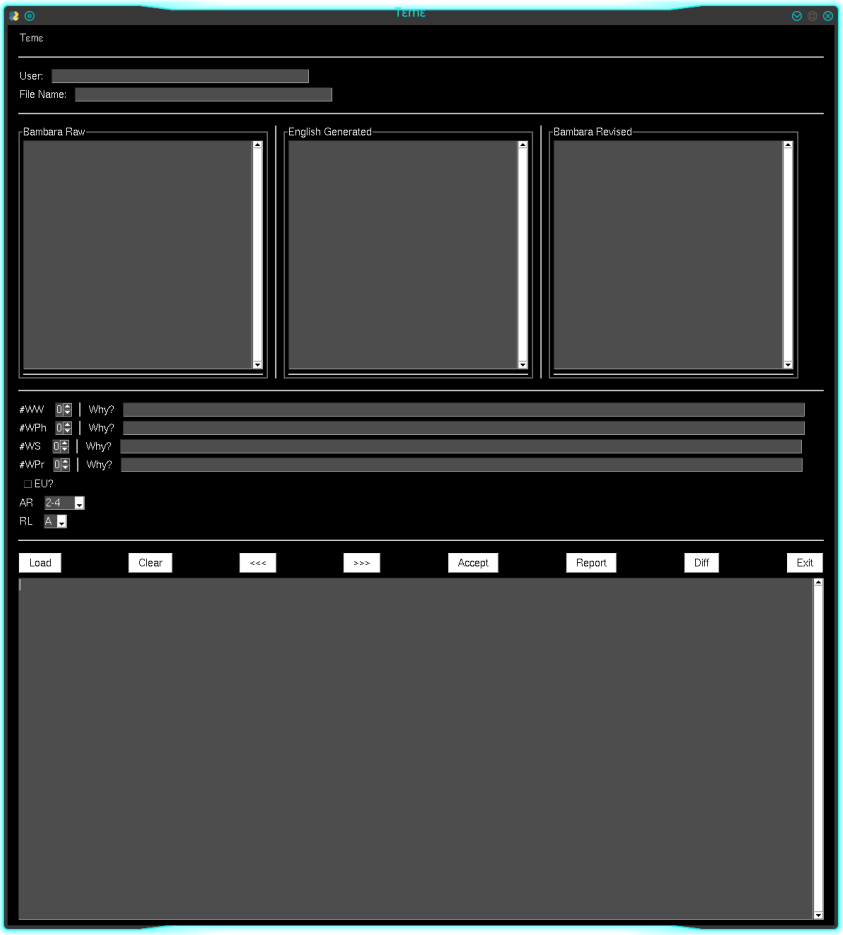

Figure 4: Our annotation tool Tεmε displays all three texts in chunks, easy to annotate. Additionally, it enables collections of diff-like stats about each story from raw translation of the generated English to the final corrected and contextualized Bambara. The first two frames "Bambara Raw" and "English Generated" are used to load both the translated Bambara story and the generated English story. Once those two are pasted in, the annotator clicks on "Load" to load all stories in their respective frames where, "Bambara Raw", "English Generated", and "Bambara Revised" display their respective and corresponding chunks of the story. The frame "Bambara Revised" is where the annotator correct the Bambara text. The annotator has the possibility to scroll through them using "< < <" for previous, and "> > >" for next. The "Clear" button clears everything. The "Accept" button records the current content of the "Bambara Revised" frame as correct. The "Report" button reports the variables of interest, which are as follows: "#WW" for the number of wrong word, "#WPh" for the number of wrong phrase, "#WS" for the number of wrong sentence, "#WPr" for the number of wrong paragraph, " EU? " for if easy to understand, "AR" for age recommendation, and "RL" for reading level. The "Diff" button generates a command to run to get a visual different between the translated Bambara text and the corrected Bambara text. Finally, the "Exit" button exits the program after confirmation.

that, in Mali, bear little resemblance to equivalent work in developed countries. Jobs like ambulatory seller and griot, integral to Malian life, had to added by the author of the book.

One strength of ChatGPT is the its ability to globally modify the language used in a text to meet requirements with respect to target reading level, grammatical and narrative complexity, imagery, and translatability. Without listing many restrictions in the prompts, ChatGPT will use idiomatic expressions and metaphors that either cannot be translated into Bambara or, if translated, don't reflect the Malian way of looking at the world or a natural use of the language. The general solution to this problem has been for the prompt engineer to request non-idiomatic, international language and to lower the age target of the generated text to produce simpler text that is more readily transformed into its equivalent in Bambara.

ChatGPT proved to be extremely powerful at the task of generating student and teacher pedagogical material related to the story once the story was in

English	He said, “Always be fair and good. Being good is the best thing.“
Bambara (Raw)	A ko: “Aw ka kɛ tilennenya ni mɔgmo ŋuman ye tuma bɛɛ. Ka kɛ ɔgmo ŋuman ye, o de ka fisa.“
Bambara (Corrected)	A ko: “Aw ka kɛ bagan ŋumanw ye ani bagan tilennenw ye. Ka kɛ ŋuman kɛ, o de kaŋi.“
English	Amina’s mother didnt like the gift from her daughter.
Bambara (Raw)	Amina ba ma diya a denmuso ka nilifɛn ye.
Bambara (Corrected)	Amina ka nilifɛn ma diya a ba ye.

Table 2: Some examples from different books (Story Number), of Google or Glosbe translation (Bambara (Raw)), the ChatGPT generated English story, the human corrected Bambara (Bambara (Corrected)), and the explanation (Insights) of why it was corrected.

final form. ChatGPT produced acceptable results approximately 90% of the time after initial prompts, usually just requiring modification of a few words here and there that come from sources other than the story itself.

A.2 Translation and Correction

Our team has been working on automatic translation of Bambara for several years, collecting an aligned bilingual machine learning dataset in Bambara and French (ref) and creating a transformer for translation (ref). In 2022, Google’s “Towards the Next 1000 languages“ (Siddhant et al., 2022) and Meta’s “No Language Left Behind“ (Costa-jussà et al., 2022) came out with multilingual machine translation models that included Bambara, giving Bambara-speakers access for the first time to relatively high quality machine translation. We took advantage of these capabilities in our project, using Google Translate (Wu et al., 2016; Johnson et al., 2017; Bapna et al., 2022) or Glosbe, which is built on top of it (Team et al., 2022).

Our authors and editors assessed that the translations produced by MT are good, but very far from being suitable for direct use in the books. A key part of our process is the generation of the source text with the knowledge that it will be translated to Bambara. We aimed to have the text, in grammatical structure and vocabulary, as close to Bambara as possible, a strategy which yielded a much higher percentage of acceptable content and made subsequent editing easier. Despite this, numerous challenges presented themselves, often having to do with the state of development of Bambara as a written language. There are very few works of literature or technical books in Bambara, or even a daily newspaper. The vast majority of native speakers of Bambara cannot read or write their mother tongue(Thiam, 2023). The editors were constantly presented with the difficult task of answering the question for texts “How can we say this in Bambara¿‘

Our experience in Bambara machine translation proved to very useful in coming up with an approach. A collaborative editing process and a tool chain of linguistic analysis applications embodied in an annotation tool, shown in Figure 4, combined with field testing and feedback provided the solution.

During annotation, similar to the story generation process, we make sure that the translation is consistent with the criteria defined. We annotate and record the variables as we go. Access to the code will be provided upon publication.

Third-Party Linguistic Tools Complementary to Tɛmɛ, to further improve and adjust the automatic translation outputs from Google Translate or Glosbe, the team used a number of tools developed by linguists for working with Bambara. These tools, Corbama (hum), Daba[12], and Bamadaba[13], all of which were developed by expert linguists at INALCO (ina), were used to standardize and structure the final language presented in the books.

Corbama Corpus Bambara de Référence, is a premier Bambara language corpus with over 11 millions words(Vydrin, 2013). This corpus was primarily used as a reference guide to address the grammatical and orthographic inconsistencies discovered post-automatic translation process.

Daba Is a morpheme analysis and semi-automatic disambiguation package (Maslinsky and

[12] https://github.com/maslinych/daba

[13] http://cormand.huma-num.fr/Bamadaba/lexicon/index.htm

Vydrin, 2019). Daba was used to assess the conformity of the target sentences to the standard Bambara, given a dictionary, and according grammatical rules.

Bamadaba is an online bilingual French-Bambara online dictionary that was highly used during the manual revision process to look up words and conduct quality assurance.

The use of these tools reinforced the necessity of human-intervention during the translation process.

A.3 Image Generation

All stories were richly illustrated with approximately 800 original images generated using a number of different tools, each with different strengths and weaknesses that were employed to find solutions to specific image generation problems. The PlaygroundAI interface with the Stable Diffusion VAE was used most often, primarily because it appeared to be best at generating simple images for storybook illustration, without excessively elaborate elements in the image composition that give images an unmistakable "AI look". Midjourney was used occasionally for its strengths in producing consistent characters and DallE3 was used when its memory enabling text instruction to progressively modify an image was useful. English language prompts often produced better results, but prompting both in English and French was successful. There were certain prompt terms that were often successful and were reused often. For example, use of the word "African" in the prompt usually resulted in an image with people of the racial type typically seen in Mali.

Every feature of the Playground interface was used at one time or another in order to obtain the desired image, including image-to-image generation, exclusion by text instruction, control traits (edge, depth, and pose) adjusting strength of prompt guidance, quality and details, and refinement, use of fixed seed, selection of different samplers, and image object and area masking, editing, erasing, and superposition. We were not able to find any combination of parameters that would consistently produce the desired results.

Many different diffusion sampling methods were used, with the most successful most often being the methods available only to "Pro" users such as DPM++2M SDE Karras. A few different style filters were used, with only a few being used consistently such as "Watercolor" but occasionally other style filters were used to obtain specific effects. For all other parameters, we typically needed to test a wide range of different combinations of values to get the desired results. While we occasionally could get a good image after only a few minutes of effort and a few dozen candidates, it was more typical to spend an hour or so per image, generating hundreds or even thousands of candidates and perfecting some images through cycles of editing and image-to-image generation. Consequently, image generation for each story typically took two to three times longer than generation of text.

Occasionally, it proved impossible to generate an image containing elements that needed to be highly accurate and were uncommon outside of the West African region. One example is the fruit called Nsaban or Zaban in Bambara, which was a central element in several stories including "Nsaban Dugu"[14]. In such cases, a solution was often found by superimposing our own photographas of the object over a generated image.

Figure 5: An example of a "bad" image. The problem here is that the people in the illustration do not have hands that resemble humans or animals hands. They have the potential to cause confusion and disturbance to children.

The greatest deficiency of the model used for the purposes of illustrating children's stories was inconsistency in representing characters and scenes. Recent improvements to Midjourney and DallE3 have improved this situation but it still remains a major obstacle when the characters and environ-

[14] https://bloomlibrary.org/RobotsMali/RobotsMali-PrintBooks-Main/RobotsMali-print-J%C9%9BkuluJ/book/doJZ8JSvrO

ments span a book's worth of images. Various creative strategies were used to cope with this, including creating characters and scenes with salient features that we would help the reader to readily associate an image with its prior manifestations, scene composition that would show persons or elements of a scene that would give the impression that the same thing was being shown from a different perspective, or simply avoiding through narrative elements to show the same thing in multiple images. Ultimately, we hoped that detectable differences in depiction of people or motifs would be accepted as a characteristic of our style of illustration. Our field tests seemed to validate this as our readers seldom pointed out inconsistencies, even as they noticed and remarked on many elements of the images.

We observed considerable bias in the generated images which required careful attention on our part. Images of both men and women appeared to us to often be highly sexualized through depiction of body type and by excessively revealing clothing. The exclusion prompts were sometimes helpful; for example, a negative prompt for "sexy" usually produced more modestly attired characters. Using "African" or "Africa" in a prompt often produced images of people living in huts surrounded by jungle and wild animals, whether appropriate or not for the setting of the story. When a modern, urban setting was needed, it was difficult to get people dressed in typical Malian clothing, though the prompt instruction "Muslim clothing" usually got a fair approximation of what was needed. Hairstyles of Malian women were also difficult to obtain, while the model readily produced African people with African hair, the model seemed to have a strong preferences for natural styles as opposed to covering the head with a scarf or braids as seen habitually in Mali.

Composing scenes with more than 2 characters, each with a set of distinctive traits was nearly impossible and even two characters proved very difficult, as the model would create images indifferently ascribing the unique characteristics of each person to one or the other, or both. Sometimes the only way to obtain a complex composition was to generate separate images and to superimpose the images.

Despite the difficulties described, we did succeed in generating images that we judged satisfactory and that strongly enhanced the reading experience of the children.

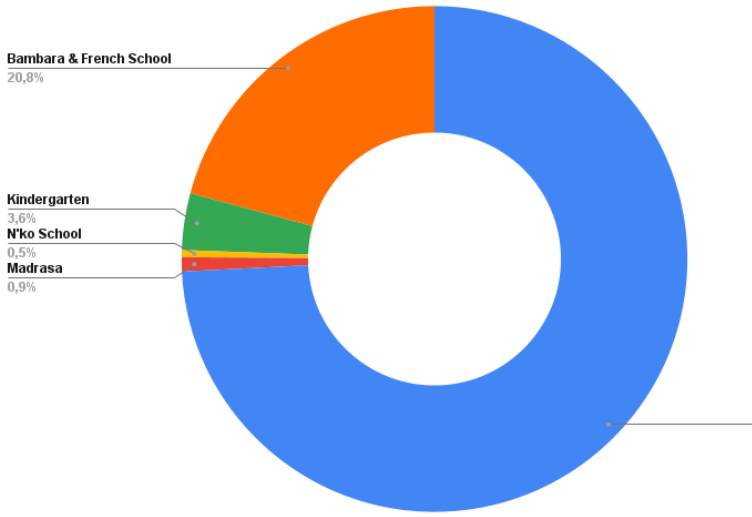

Figure 6: The different backgrounds of languages spoken and known to children with whom we did the field tests.

A.4 Perspectives

Perspectives of parents and teachers in urban versus rural areas.

Questions Urban Sampling	Agree	Disagree
Do you think Bambara should be incorporated into school curricula?	52%	48%
Do you want your children to learn to read Bambara?	22%	78%
Do you anticipate that this project can have a positive impact?	17%	83%

Table 3: Parent and Teacher Perspectives (Sample Size: 23)

Questions Rural Sampling	Agree	Disagree
Do you think Bambara should be incorporated into school curricula?	100%	0%
Do you want your children to learn to read Bambara?	100%	0%
Do you anticipate that this project can have a positive impact?	100%	0%

Table 4: Parent and Teacher Perspectives (Sample Size: 15)