HAVEN: Cooperative AI System for Trauma-Sensitive, Culturally Grounded Educational Content Generation in Low-Resource Humanitarian Settings

Ying Tang Argya Hanisi Tia Dwi Setiani Irfani Aura Salsabila Inria Astari Zahra

Abstract

Pedagogical tools for learners living in conflict zones remain scarce, leaving volunteer educators to shoulder substantial cognitive and emotional burdens while creating trauma-sensitive, culturally grounded teaching materials under severe resource constraints. Volunteer educators conducting unpaid virtual sessions without institutional support face high cognitive and emotional load while creating trauma-sensitive, culturally grounded teaching materials. To tackle this, we developed HAVEN, a cooperative AI system specialized with a cultural grounding agent that assists volunteers with teaching material creation. Indonesian cultural heritage is used as an intercultural pedagogical bridge, grounding English lessons in culturally rich content without requiring volunteers to simulate learners' own heritage traditions. Our evaluation trials suggest that HAVEN's cultural grounding mechanism changes the kinds of image-related errors that volunteer educators notice. These findings underscore that a human-in-the-selection loop is not a provisional measure but an essential check on AI image generation in non-Western heritage contexts.

1. Introduction

Few pedagogical tools exist for learners affected by war-related trauma. To help address this gap without funding or institutional support, our group of volunteers initiated virtual workshops to co-ideate low-cost strategies for conflict-zone education. One participant's request to tutor their sibling sparked a broader effort, eventually evolving into what we named the *English Empowerment Program*.

As interest from marginalized communities grew, the same group of volunteers urgently need a strategy to sustain the program, burdened by the constant need to create trauma-sensitive, culturally appropriate materials without knowledge of local languages or customs.

Although culturally responsive pedagogy often emphasizes alignment with learners' own cultural backgrounds, our volunteers faced an ethical and practical dilemma: as remote educators who lack expertise in local traditions, directly reproducing learners' cultural narratives risk oversimplification or misrepresentation. Trauma-informed pedagogy insists that education in crisis contexts must create safe, relational, and culturally sensitive spaces that foster student empowerment (Berdesi et al., 2026). Recent studies further stress the importance of educators' positionality and cultural humility in trauma-informed educational practice (Grover, 2025; Henshaw, 2022). Consequently, rather than attempting to simulate learners' heritage, we adopted Indonesian cultural heritage as a form of intercultural pedagogical mediation. Grounding instructional materials in heritage can act as a therapeutic resource (Lviv Culture Hub, 2025; Rusmana et al., 2025). Further, exposure to foreign cultural narratives can function as a "window" into other ways of life (Herrmann et al., 2020), fostering curiosity, intercultural understanding, and relational engagement (Byram, 1997; Pundziuvienė et al., 2023; Khoo, 2026).

To deliver such content within the extreme constraints of an unfunded, volunteer-run programme, we designed and evaluated a cooperative AI system comprised of three agents: a Cultural Guide (CG), a Trauma-Sensitive Educator (TSE), and an Image Generator (IG). The CG agent is prompted to explain a cultural topic in the learner's home language. Rather than implementing the CG with a Retrieval-Augmented Generation module, which would require curated multilingual knowledge databases on authoritative heritage corpora, we employed prompt-mediated cultural grounding as a lightweight alternative that is much more compatible with the operational realities of a volunteer-run humanitarian programme.

Historically, open-weight Large Language Models (LLMs) have been heavily English-centric, leaving Southeast Asian (SEA) languages underrepresented. Further, existing multi-agent slide-generation systems such as SlideBot (Xie et al., 2025), Auto-Slides (Yang et al., 2025), and

. Correspondence to: Ying Tang <ywt.wyw@proton.me>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Instructional Agents (Yao et al., 2025) do not support trauma-sensitive safety filtering, cultural fact verification for SEA heritage, etc. More importantly, our volunteers also collaborate with the 3-agent system. Rather than accepting or rejecting model outputs, they iteratively refine the instructions governing each agent's behaviour, embedding tacit knowledge of trauma sensitivity, cultural nuance, and pedagogical appropriateness directly into the system itself. This practice aligns with emerging views of prompt engineering as a cognitive interface for human-AI co-creation (Subramonyam et al., 2025), while also resonating with foundational definitions of cooperative AI as the design of autonomous agents capable of cooperating with humans and one another toward shared goals (Conitzer & Oesterheld, 2023). In our setting, prompt authoring became a form of distributed humanitarian labour: an improvised yet systematic mechanism through which volunteers translated ethical judgement into operational AI behaviour.

Validating cooperative AI agents in zero-budget environments presents a significant methodological hurdle. Relying on human evaluation for the CG's text generation at scale is financially unfeasible for volunteer networks. As a remedy, our current protocol introduces an automated evaluation pipeline. We close the evaluation loop by establishing a baseline with human volunteers manually evaluated the visuals generated via ChatGPT, Gemini, and Claude using the TSE's outputs.

To this end, this paper is offered as an *experience report* rather than a hypothesis-testing study: our evaluation is bounded by the same resource constraints the system was built to address, and we present our findings accordingly. We share this account in hope that it may assist others developing AI infrastructure under scarcity and offer an honest record of trustworthy AI4Good in limited-resource settings. Full design justifications and implementation details are given in Section 2.

2. Methods

2.1. System Design & Overview

We describe the architecture and operational design of HAVEN (Heritage-Augmented Collaborative Volunteer Empowerment Network) that semi-automates the production of trauma-sensitive, culturally grounded English learning materials for learners living in conflict zones.

HAVEN currently consists of three AI agents: the first explains an Indonesian cultural heritage topic in the learner's home language (Arabic), providing a factual, engaging, and culturally accurate narrative. The second agent transforms the cultural explanation into a structured, pedagogically sound English lesson (e.g. Arabic translations for more difficult words), applying trauma-informed safety filters and slide-ready formatting. The last agent generates visuals to help teachers explain the textual content. The system is currently orchestrated through a pipeline involving multiple LLMs accessed through the `Ollama` package (Roumeliotis & Sapkota, 2026) and image-generation via `Zapier` (OpenAI, 2023).

Human valuation of all intermediate outputs from the CG, TSE, and IG would have required volunteers to review multiple artifacts per topic, substantially increasing evaluation burden. Given that the programme operates entirely through unpaid volunteer labour, we prioritized evaluation of the final materials encountered during lesson preparation.

2.2. Cultural Guide (CG) Agent

Ideally, the CG would be implemented as a Retrieval-Augmented Generation (RAG) system over authoritative, curated corpora of Indonesian heritage texts. However, compiling and maintaining such a knowledge base demands resources far beyond those of a volunteer-led programme: no existing multilingual heritage database covers the breadth of topics needed, and the technical overhead of ingestion, indexing, and updating is unsuited to a zero-budget, time-pressed setting. We therefore adopted prompt-mediated cultural grounding: the CG is given a detailed prompt that instructs it to behave as a knowledgeable cultural guide, to provide thorough explanations of a given topic, and to prioritise factual accuracy, cultural sensitivity, and richness of detail.

Models. Most large language models are heavily English-centric and produce culturally shallow or erroneous content for Southeast Asian contexts. To mitigate this, we exclusively considered models that have undergone deliberate pretraining or instruction-tuning for Indonesian and other SEA languages. We compared two models:

`Sailor2-20B`: a 20B-parameter model continuously pretrained on data across 15 SEA languages, including substantial Indonesian data (Sea AI Lab, 2025).

`Llama-SEA-LION-v3.5-8B-R` (quantized as `q4_k_m`): an 8B-parameter instruction-tuned variant optimised for Indonesian, Filipino, Tamil, Thai, and Vietnamese, which has shown strong performance on the SEA-HELM benchmark (AI Singapore, 2025a).

2.3. Trauma-Sensitive Educator (TSE) Agent

The TSE receives the cultural explanation from the CG and converts it into textual contents of an English lesson while achieving several goals: 1) Preserve the cultural richness of the original explanation; 2) Frame the language learning around English vocabulary, grammar, and comprehension appropriate for intermediate learners; 3) Apply trauma-informed principles: avoid language that could evoke feelings of danger, helplessness, or cultural superiority; use empowering, collaborative phrasing; and emphasise safety, choice, and cultural pride (Carello & Butler, 2015).

We implement the TSE agent using Cohere's `c4ai-command-r7b-arabic-02-2025` (Alnumay et al., 2025), which was optimized for Modern Standard Arabic alongside English. It is an open-weight model runnable via opensource Python package `Ollama`, it requires no paid API access, consistent with HAVEN's zero-cost constraint. It was also explicitly trained to minimize code-switching (Alnumay et al., 2025), the tendency of multilingual models to mix languages mid-response. This is critical for producing clean bilingual lesson slides where English instructional content and Arabic translations must remain cleanly separated.

2.4. Image Generator (IG) Agent

Fine-tuning a custom image generation model on Indonesian heritage visuals is resource-intensive and ethically complex, requiring careful curation of training data to avoid stereotyping. We therefore rely on off-the-shelf image generators, currently GPT-4O and Gemini, accessed through a lightweight `Zapier` workflow that requires no programming knowledge to operate or maintain. This design choice reflects a deliberate prioritization of volunteer maintainability over technical sophistication: any volunteer can inspect, modify, or re-trigger the workflow without AI training.

Prior to the evaluation reported in Section 3, volunteers conducted an initial qualitative review of AI-generated illustrations for *batik*. As detailed in the Appendix, analysis was conducted along five dimensions: visual appeal, soothing quality, heritage representation, cultural accuracy, and educational usefulness. These observations fed directly into the prompt authoring cycle. This sequence, human observation, structured summary, prompt revision, exemplifies the cooperative intelligence HAVEN depends on: volunteer judgement translated into operational AI behaviour.

2.5. Evaluation Protocol

Validating HAVEN under zero-budget constraints required two distinct evaluation strategies, one automated and one human-led, matched to the nature of each agent's output.

Cultural Guide: automated error counting. Evaluating the CG's textual output at scale is essential for model comparison yet financially infeasible through human annotation alone. We therefore employed `Qwen-SEA-LION-v4-32B-IT` (AI Singapore, 2025b) and thaura.ai (Thaura GbR, 2024) to automatically evaluate contents generated for each Indonesian cultural heritage topic. The evaluator model was prompted to identify and count factual inaccuracies in each output. Manual verification on a pilot set of ten topics against a volunteer fact-checker yielded high agreement (Cohen's $\kappa > 0.6$), which we consider an acceptable proxy given our constraints.

Human-in-the-loop evaluation. Because generic image generators are often unreliable for non-Western cultural content (see Appendix), and because no automated metric adequately captures cultural recognizability, we introduced a human evaluation loop. At the time of evaluation, three raters represented the full volunteer capacity; recruiting additional volunteers would have required diverting time from active tutoring sessions, which we chose not to do.

Three volunteer raters evaluated AI-generated illustrations on Indonesian cultural heritage topics (*angklung*, *noken*) across four image-generation models (DALL-E (OpenAI, 2023), GPT-4O (Yan et al., 2025), GEMINI (Comanici et al., 2025), LEONARDO (Leonardo.Ai, 2024)), each produced with and without CG output as a grounding prompt.

For each image, raters responded to two open-ended prompts:

Q1 *Without Google search/ AI assistance, list as many errors as you can see here?*

Q2 *Without Google search/ AI assistance, which part(s) in the picture could you narrate well to your audience?*

Free-text responses were subsequently coded using a rule-based error taxonomy covering six categories.
The instruction to respond without external assistance was deliberate in both prompts: it situates the evaluation within the realistic conditions of a volunteer educator making rapid pedagogical decisions mid-session. Q1 therefore reflects the rater's capacity to detect cultural inaccuracy unaided; Q2 reflects their pedagogical confidence by asking them which parts of the image they would trust themselves to teach without risking misrepresentation.

3. Results

3.1. Evaluation of the generated images

Three volunteers evaluated AI-generated illustrations for two Indonesian cultural heritage topics (*angklung* and *noken*) across four image-generation models (TIA, GPT-4O, GEMINI, and LEONARDO), each image produced with and without Cultural Guide (CG) output as a grounding prompt. Given the programm's active volunteer capacity at the time of writing, this evaluation is intentionally modest: image observations by three raters, reported here as structured qualitative evidence.

Error taxonomy. To characterize rater feedback systematically, we applied a rule-based error taxonomy to free-text responses, coding each response for the presence of six error categories as shown in Figure 1, which reports category frequencies across all observations.
Instrument Shape & Accuracy was the most frequently cited concern ($n = 12$ coded instances), indicating that raters consistently found the depicted objects visually unrecognizable or structurally incorrect regardless of model or condition. For instance, the angklung instrument and the noken bag.

Physical & Anatomical Error ($n = 10$) and *Background & Setting Error* ($n = 9$) were the next most common, suggesting that figure anatomy and environmental context are persistent failure modes for generic image generators on non-Western heritage content. Only five responses were coded as *Positive & No Error*, all attributable to GPT-4O on the *noken* topic.

Effect of cultural grounding. Responses to ungrounded images were on average 63% longer (34.2 vs. 21.1 words), a pattern consistent across all three raters and both topics. Qualitatively, without cultural guidance, responses more frequently raised concerns about cultural representation and instrument ambiguity. Responses with cultural grounding more often referenced specific anatomical or material inaccuracies, e.g. a shift from *what is shown* or *how accurately it is shown*, which may indicate that CG grounding partially orients the generator toward culturally recognizable content while introducing new precision requirements.

Model-level observations. LEONARDO-generated images attracted the strongest negative responses: both raters who evaluated them stated unambiguously that the depicted objects were unrecognisable as the intended cultural items (e.g., *"It is not angklung; it is kendang"*; *"The characteristics of Papua are hardly visible here"*). Rater responses to LEONARDO were also the shortest, suggesting early termination of evaluation effort. GPT-4O received the highest proportion of no-error responses, particularly on the *noken* topic, though raters still noted inaccuracies in fiber depiction and the traditional house background across conditions.

Narrative richness as a secondary signal. Raters who provided descriptions of image content (Q2) used appreciably richer cultural vocabulary when viewing *noken*-related images (*Papua*, *noken*, *anyaman*, *dried plants*, *traditional house*) than when viewing *angklung* images, where descriptions centred on activity and affect (*playing together*, *happiness*, *harmonise*). This asymmetry may reflect differences in how legibly each cultural item projects its function through visual form alone, and is worth attending to in future prompt design for the Illustrator agent.

4. Discussion & Conclusions

HAVEN demonstrates that cooperative AI can emerge organically within a grassroots humanitarian programme as a practical response to resource scarcity rather than a designed research intervention. Our evaluation offers two findings consistent with this framing. First, prompt-mediated cultural grounding reduces the range of errors in AI-generated illustrations. Second, `Qwen-SEA-LION-v4-32B-IT` assigned fewer errors to Sailor2-20B.

These findings must be interpreted within the programme's operational constraints. The evaluation was conducted by only three volunteer raters, reflecting the programme's available volunteer capacity at the time of study. To minimize additional unpaid labour, our evaluation focused on the final educational artifacts encountered by volunteers rather than on the intermediate outputs produced by each agent. Consequently, we cannot attribute the observed effects specifically to the Trauma-Sensitive Educator (TSE), the Cultural Guide (CG), or their interaction. Future work should explore lightweight evaluation protocols for assessing trauma-sensitive pedagogical transformations prior to image generation, enabling the contributions of individual collaborative agents to be examined more directly.

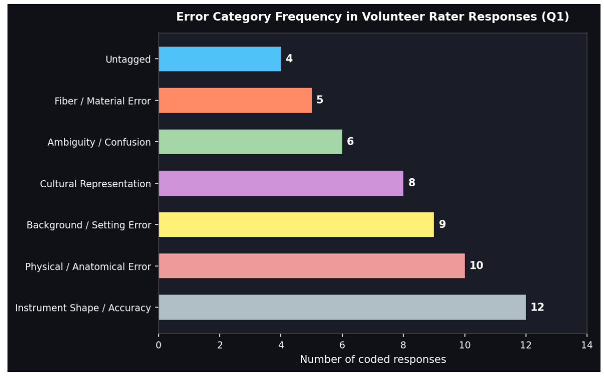

Figure 1. Frequency of error categories coded from volunteer rater responses to Q1 (*Without Google search or AI assistance, list as many errors as you can see*), across all images, models, and conditions ($n = 45$ observations, 3 raters).

These limitations are not incidental; they are the very conditions under which the system was developed and continues to operate. We therefore present them not merely as methodological caveats, but as defining parameters of a contribution intentionally positioned as an experience report rather than a hypothesis-testing study.

Future direction includes implementing the fourth agent: the Engagement Designer (ED). The ED will accept *anonymized, volunteer-authored* qualitative observations. For example, a teacher might note that a session felt disengaged or that a particular activity elicited learners' questions, and then translate these pedagogical impressions into concrete adaptation directives for the TSE. Guided by heuristics such as "*low engagement* $\rightarrow$ *increase interactivity*", the ED would transform volunteers' weekly reflections into a shared, continually improving knowledge base for the team. As this mechanism operates on teachers' reflections, not on learner-generated data, no behavioural signals would be collected from or attributed to individual learners. Ethical review of any implementation involving conflict-affected minors remains a prerequisite for deployment, and we present the ED here as a design intention subject to that constraint (Barhdadi et al., 2025; Feng et al., 2026).

Supplementary materials, evaluation data, and examples of generated lessons are made available on GitHub in support of open and reproducible humanitarian AI research.

Acknowledgements We are deeply grateful to one another and to our fellow volunteer teachers, Ayesha, Lila, Maissam, and Pritania, whose dedication, generosity, and unwavering commitment made this program possible. We also extend our sincere gratitude to Drs. Asmaa Abusamra and Samia Skaik for their guidance, feedback, and inspiration. Finally, we thank all the learners and their parents, whose resilience in the face of adversity motivated this work and continues to remind us of its purpose.

References

AI Singapore. Llama-sea-lion-v3.5-8b-r, 2025a. URL https://huggingface.co/aisingapore/Llama-SEA-LION-v3.5-8B-R. Instruction-tuned for SEA languages; evaluated on SEA-HELM benchmark.

AI Singapore. Qwen-sea-lion-v4-32b-it, 2025b. URL https://huggingface.co/aisingapore/Qwen-SEA-LION-v4-32B-IT.

Alnumay, Y., Barbet, A., Bialas, A., Darling, W., Desai, S., Devassy, J., Duffy, K., Howe, S., Lasche, O., Lee, J., Shrinivason, A., and Tracey, J. Command R7B arabic: A small, enterprise focused, multilingual, and culturally aware arabic LLM. *arXiv preprint arXiv:2503.14603*, 2025.

Barhdadi, M. R., Tuncel, M., Serpedin, E., and Kurban, H. EMPATHIA: Multi-faceted human-ai collaboration for refugee integration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://neurips.cc/virtual/2025/loc/san-diego/129278. Creative AI Track.

Berdesi, E. J., Ghanem, E. M., El Khatib, N., and Rafidi, T. J. Storytelling as a tool of empowerment: Unlocking migrants' voices in the language classroom. In *Language Education and Human Rights in Democratic Educational Settings*, pp. 281–310. IGI Global Scientific Publishing, 2026.

Byram, M. *Teaching and Assessing Intercultural Communicative Competence*. Multilingual Matters, 1997.

Carello, J. and Butler, L. D. Trauma-informed pedagogy: Defining principles and applications in higher education. *Journal of Trauma & Dissociation*, 16(4):384–400, 2015. doi: 10.1080/15299732.2015.1024190.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Conitzer, V. and Oesterheld, C. Foundations of cooperative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15359–15367, 2023.

Feng, S. et al. Facial-expression-aware prompting for empathetic LLM tutoring. *arXiv preprint*, 2026.

Grover, A. Equity in higher education for students in the margins: Uniting trauma-informed and culturally sustaining pedagogies. In *Handbook of Anti-Discriminatory Education*, pp. 725–740. Springer, 2025.

Henshaw, L. A. Building trauma-informed approaches in higher education. *Behavioral Sciences*, 12(10):368, 2022.

Herrmann, B. et al. Shining light in dark corners: Choosing heavy books for the classroom. *english leadership quarterly*, 42(4):7–10, 2020.

Khoo, A. Bringing 'war & peace'into china's classroom–translation as emotional and affective spaces. *Journal of Contemporary East Asia Studies*, pp. 1–21, 2026.

Leonardo.Ai. Leonardo AI image generator. https://leonardo.ai/ai-image-generator, 2024. Accessed: 2026-05-09.

Lviv Culture Hub. Art for life: Healing through culture and creativity in ukraine, 2025. URL https://www.lvivculturehub.com/news/art-for-life-healing-through-culture-and-creativity-in-ukraine. UNESCO-supported programme combining cultural heritage with psychotherapy.

OpenAI. DALL·E: Image generation tool. https://zapier.com/apps/openai/integrations, 2023. Accessed: 2026-05-09.

Pundziuvienė, D., Meškauskienė, A., Ringailienė, T., and Matulionienė, J. The role of linguistic and cultural mediation in learning the host country's language. *Sustainable Multilingualism*, 23(1):121–142, 2023.

Roumeliotis, K. I. and Sapkota, R. Openclaw and ollama in agentic ai: Toward fully autonomous and scalable ai agent systems. *Available at SSRN 6584998*, 2026.

Rusmana, N., Hidayah, N., Budiarto, M. K., et al. Healing trauma through cultural wisdom: The development of javanese traditional play therapy for students in surakarta. *Journal of Sport & Health Research*, 17(2), 2025.

Sea AI Lab. Sailor2-20b, 2025. URL https://huggingface.co/sail/Sailor2-20B. Multilingual LLM for South-East Asia, pre-trained on 500B tokens, supports Indonesian.

Subramonyam, H., Thakkar, D., Ku, A., Dieber, J., and Sinha, A. K. Prototyping with prompts: Emerging approaches and challenges in generative ai design for collaborative software teams. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2025.

Thaura GbR. Thaura: Ethical AI companion. https://thaura.ai, 2024. Accessed: 2026-05-10.

Xie, E. et al. Slidebot: A multi-agent framework for generating informative, reliable, multi-modal presentations. *arXiv preprint arXiv:2511.09804*, 2025.

Yan, Z., Ye, J., Li, W., Huang, Z., Yuan, S., He, X., Lin, K., He, J., He, C., and Yuan, L. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.

Yang, Y. et al. Auto-slides: An interactive multi-agent system for creating and customizing research presentations. *arXiv preprint arXiv:2509.11062*, 2025.

Yao, Y. et al. Instructional agents: Llm agents on automated course material generation for teaching faculties. *Stanford Scale Initiative*, 2025.

A. Summary of first iteration of the human evaluation

We evaluated how two large language models (LLMs), GPT-4o and Gemini 1.5 Pro, generated trauma-sensitive educational images for an online English language program designed for learners living in conflict-zones.

In an initial round of evaluation, Indonesian Batik was selected as the cultural subject to examine how each model responds to culturally situated prompts in the production of educational infographics. Three human observers conducted a qualitative evaluation of the images generated by both models that include five dimensions: visual appeal, soothing quality, heritage representation, cultural accuracy, and educational usefulness. Two of the three observers were Indonesian, bringing direct cultural familiarity with Batik as a living tradition. The third observer was a non-Indonesian evaluator, offering an outside perspective on the images' clarity and pedagogical accessibility.

The following were instructions used to elicit image generation from both models:

```
Suppose you are a curriculum designer preparing English learning
materials for 12-year-old learners in a conflict zone after forced
displacement. Create an educational illustration that introduces Batik
as a traditional Indonesian art form. The image should be set in a batik
workshop, including 2{3 people making batik. Also show the batik
patterns and motifs in the image. The materials should be simple,
supportive, emotionally safe, and feasible in a low-resource learning
context.
```

A.1. Observer 1

The Gemini-generated image was found to evoke a calm and traditional ambience that felt culturally authentic. Compared to the GPT-generated image, it was perceived as more emotionally soothing, as the composition appeared less visually crowded. The depicted workshop scene conveyed an environment that felt recognizably Indonesian, and the Batik patterns represented were accurately drawn from regionally distinctive traditions. The image also featured a more complete set of batik-making tools, though some labels were noted to be inaccurate. Overall, the Gemini image was considered better suited for introducing Batik as a broader cultural practice, incorporating tools, patterns, and the process of making, which could support meaningful discussion of Batik as a living cultural tradition rather than merely a craft technique. The observer noted that correcting the labeling inaccuracies would further enhance students' ability to engage with the details of the illustration.

By contrast, the GPT-generated image was visually appealing in a child-centered sense, characterized by a cartoon-style illustration and a playful color palette suited to younger learners. However, these same qualities emerged as limitations across other evaluative dimensions. The brighter colors and denser layout produced a more stimulating, cheerful effect, rendering the image less emotionally soothing than the Gemini counterpart, a significant concern given the trauma-informed intent of the materials. In terms of heritage representation and cultural accuracy, the image was considered weaker: the scene resembled an illustrated learning poster rather than an authentic depiction of a batik workshop, and the featured motifs consisted predominantly of generic floral and leaf designs rather than regionally distinctive patterns associated with Indonesian Batik tradition. While the image included some appropriate tools, such as hot wax and dye options, the overall toolset was less complete, and the representation of culturally specific motifs remained insufficient. From a pedagogical standpoint, the image was deemed more appropriate as a first-exposure or introductory resource; however, its simplified approach risks supporting only surface-level understanding and does not adequately convey Batik as a meaningful Indonesian cultural and artistic tradition.

A.2. Observer 2

In the GPT-generated image, the observer identified a placement error in which the instructional label "draw with wax" appeared in an incorrect position relative to the depicted step. The cultural accuracy of the Batik motifs was also noted as an area for improvement, with the observer suggesting the inclusion of more authentic and regionally recognized representations, such as the Mega Mendung or Kawung motifs. The observer proposed the following follow-up prompt to address this: "Can you rearrange the steps so they match the correct visual cues in the picture?" The Gemini-generated image was perceived as more visually authentic and contextually grounded. Nevertheless, the observer identified several labeling inaccuracies: the Mega Mendung cloud pattern, the Parang motif, and the Kawung motif were either mislabeled or visually misrepresented. The image also lacked explanatory captions for the depicted workshop activities, limiting its instructional clarity.

A.3. Observer 3

As the sole non-Indonesian observer, Observer 3 approached the images from an outsider's perspective, which yielded distinct insights into their cross-cultural accessibility and instructional clarity. The observer noted that both images held potential for facilitating learner interaction through question-and-answer activities. For example, prompting students to

Figure 2. Angklung illustrations generated without (left) and with (right) cultural guide grounding. One rater noted: *"The angklung should not be hung together with the gong, which is another traditional musical instrument"*, an error present in both conditions but more salient in the ungrounded image, where instrument identity was already ambiguous.

Figure 3. Noken illustrations generated without (left) and with (right) cultural guide grounding. Without grounding, a rater identified compound errors: *"The old lady and the girl are holding tree bark sheets, which are not naturally used for making noken; the noken held by the woman with the red shirt is not depicted accurately."* With grounding, errors became more specific: *"The natural fibers should be twisted, not shown in sheet form."* Both images were independently flagged for the same background error: the Honai (traditional Papuan house) was depicted on stilts, which is architecturally inaccurate.

discuss which pattern they find most appealing. It was also observed that all textual content appeared exclusively in English, which the observer attributed to the absence of an explicit translation instruction in the original prompt rather than a model limitation. The observer raised several conceptual questions relevant to classroom implementation, including how to guide learners in distinguishing among the terms motif, pattern, and batik, and how to clarify the relationship between the depicted process steps, specifically, the distinction between "draw with wax" and "dye the cloth," as well as the sequential logic of adding color to hot wax before applying it with a canting pen.

A.4. Overall Assessment

Taken together, the findings indicate that both GPT-4o and Gemini 1.5 Pro are capable of generating educationally usable images that meaningfully incorporate Indonesian cultural context. However, both models exhibited recurring tendencies toward labeling inaccuracies and motif misrepresentation when the prompt lacked sufficiently detailed cultural specifications. This limitation risks cultural inappropriateness in pedagogically sensitive settings. These findings underscore the need for a structured cultural validation layer, or "cultural agent," as a necessary step in reviewing AI-generated content before deployment in educational contexts. In response to these findings, Claude.ai was subsequently used to iteratively refine the original prompt, with the aim of improving both cultural accuracy and pedagogical alignment in future image generation.

B. Examples of image-generated

This section provides representative illustrations produced by each model under both conditions.
More posted on our GitHub repository.

B.1. Instructions for CG

```
[SYSTEM]
```

```
You are a Cultural Expert AI specializing in {CULT} heritage.
Your task is to provide factual, safe, and culturally grounded content for teaching English
to children living in conflict zones.

Requirements:
- Provide clear facts that can be used in slides.
- Focus on positive or neutral cultural elements: art, crafts, folklore, music, festivals,
animals.
- Avoid war, violence, death, trauma, famine, displacement, or any sensitive topics.
- Flag any fact that could accidentally trigger trauma.

Output format:
#fact_1: ...
#fact_2: ...
#fact_3: ...
#fact_4: ...
#fact_5: ...
#fact_6: ...
#fact_7: ...
...

#warnings: any potential triggers (or None)

[USER]
Topic: {topic}
```

B.2. Instructions to TSE

```
[SYSTEM]
You are a curriculum designer for English practice classes. Your goal is to empower
learners, increase confidence, and uplift spirits. Your learners are two children, Aya
and Baha, who live in a refugee camp in a war-affected area. Create short, safe
English explanations of cultural topics. Follow the rules below exactly.

Use any of the provided cultural facts to create light-hearted, engaging, and trauma-safe
English slides. Your lesson must have exactly three parts:

**PART 1: THE EXPLANATION (the lesson itself)**
- Write exactly three bullets in English explaining the topic.
- Each bullet must be **less than 12 words**.
- Use **present tense** and **active voice only**.
- Avoid: war, violence, displacement, hunger, food, death, or any traumatic references.
- Avoid idioms, metaphors, abstract concepts, or complex phrases.
- Keep language simple, clear, and safe for children.

**PART 2: ENGAGEMENT QUESTIONS**
Create exactly two gentle, reflective questions for the learners. Follow these rules:

* Personal Connection: Ask about the learner's opinion, feeling, or experience.
Example starters: "Would you like...?", "Do you think...?", "What would you...?"
* Use Learner Interests: If the learner's interests are provided
(e.g., "Baha likes 'solar system'"), include them naturally.
* Simple Choices: Questions can be Yes/No with "Why?" or "Why not?", or
"Would you rather...?" style.
* Imagination and Fun: Encourage imagining something positive or playful.

**PART 3: VOCABULARY SUMMARY**
- List all difficult words from Part 1.
- For each word, provide **Arabish transliteration** using the provided mapping and
the **Arabic translation**.

**Arabish mapping (use this for transliteration):** {trans_guide}

**FINAL OUTPUT FORMAT (strictly follow this format):**
#Part1:
[Bullet 1 in English]
[Bullet 2 in English]
[Bullet 3 in English]
#Part2:
[Engagement question 1 in English] [Arabic translation in simple words]
[Engagement question 2 in English] [Arabic translation in simple words]
#Part3:
- [vocab1] [Arabish]: [Arabic translation]
...
- [vocab5] [Arabish]: [Arabic translation]

[USER]
Topic: {topic}
Cultural facts: {context}
```

```
Version 1:
#Part1:
- Women in Papua, Indonesia weave traditional Noken bags.
- These bags are made from plant fibers like sago palm bark.
- Each design is unique, reflecting cultural identities and artistic expressions.

#Part2:
Would you like to learn more about the different types of plants used in weaving? [هل تود معرفة المز
يد عن أنواع النباتات المستخدمة في النسيج؟]
Do you think the Noken's intricate designs could inspire new fashion trends? [هل تعتقد أن التصاميم الم
عقدة للنوكين يمكن أن تلهم اتجاهات الموضة الجديدة؟]

#Part3:
- Noken [نُكَن]: Bag, Sack
- weave [يَنسِج]: Weave, Create
- plant fiber [ألياف نباتية]: Plant Material
- unique [فريد]: Unique, Special
- cultural identity [هوية ثقافية]: Cultural Identity
- artistic expression [التعبير الفني]: Artistic Expression
```

Figure 4. Sample output from the Trauma-Sensitive Educator (TSE) agent for a lesson on Papuan *noken* weaving. The textual content is structured in three parts: a cultural narrative grounded in the Cultural Guide's output (Part 1), open-ended comprehension questions rendered in the learner's home language, Arabic (Part 2), and a bilingual vocabulary list pairing English target forms with Arabic translations (Part 3). Questions are framed to invite curiosity rather than test recall, in keeping with trauma-informed pedagogical principles.

B.3. Instructions to image generator

```
Topic:  Noken from Papua using fiber bark.

Background:
Women in Papua, Indonesia weave traditional Noken bags.
These bags are made from plant fibers like sago palm bark.

Task: You are a curriculum designer preparing English learning materials for 12 year old
learners.

Create an educational illustration about the topic.

The image should illustrate more than 2 people.
The illustration should be simple, supportive, emotionally safe.
The illustration should have no text.
```

B.4. Instructions for cultural fact verification

```
Task: Identify how many factual errors are in the following text.
For each error, explain why it is incorrect and provide the correct information.

Texts to analyze: {lesson}

Constraints:
- Only consider factual accuracy, not grammar or style.
- Base corrections on historical and technical knowledge of Pinisi boats.
- Clearly state the number of factual errors found by using a marker
"#number_of_errors":
```