

## Evaluation of Gen-AI (LLM) Results on Prompts in the Batak Language

Ade Siahaan<sup>1\*</sup>, Angelina Nadeak<sup>1</sup>, Jeremy Samosir<sup>1</sup>, Jesica Siburian<sup>1</sup>, Boy Hutajulu<sup>2</sup>

<sup>1</sup>Information System, Faculty of Informatics and Electrical Engineering,  
Institut Teknologi Del, Toba, North Sumatera, 22381, Indonesia

<sup>2</sup>Information System, Faculty of Informatics,  
Universitas Mikroskil, Kota Medan, North Sumatera, 20212, Indonesia

\*Corresponding Author: [iss22029@del.ac.id](mailto:iss22029@del.ac.id)

### Article Info

Received: 2 August 2026

Revised: 6 September 2026

Accepted: 9 September 2026

Available online: 30 September 2026

### Keywords

Batak language, CRISP-DM,  
generative AI, large language  
models, low-resource NLP,  
multilingual evaluation.

### Abstract

While Large Language Models (LLMs) demonstrate remarkable NLP capabilities, their performance on low-resource regional languages like Toba Batak remains underexplored. This study presents a rigorous empirical evaluation of four LLMs—ChatGPT, Gemini, Perplexity, and the domain-adapted Toba-LLM—on prompts written in the Toba Batak language using the CRISP-DM framework. Using a benchmark dataset spanning five task domains (information-seeking, instruction-following, creative generation, role-playing, and text-based tasks), model outputs were evaluated across six metrics: lexical similarity, keyword overlap, length ratio, multilingual cosine similarity, cultural relevance, and fluency. Experimental results reveal that general-purpose models outperform Toba-LLM in semantic similarity (cosine similarity: ChatGPT 0.628, Gemini 0.599) and fluency (ChatGPT 0.968, Gemini 0.945), whereas Toba-LLM demonstrates superior lexical alignment (0.0597) and conciseness (length ratio of 2.895). Nevertheless, cultural relevance remains a major challenge across all models (scores between 0.326 and 0.559), with qualitative findings highlighting frequent language inconsistency and Indonesian dominance bias. These findings emphasize the necessity of culturally grounded metrics, enriched regional corpora, and language-consistency constraints to advance LLM applicability in digital humanities and regional language preservation.

## 1. Introduction

Large Language Models (LLMs) represent a transformative breakthrough in artificial intelligence, utilizing deep neural architectures with billions of parameters pre-trained on massive text corpora to capture statistical, syntactic, and semantic regularities of human language [1], [2]. The emergence of state-of-the-art models—including OpenAI's GPT-4 (ChatGPT), Google's Gemini, and retrieval-augmented systems such as Perplexity—has revolutionized information retrieval, automated reasoning, and cross-lingual communication. High-performing language models are characterized by four foundational capabilities [1], [2]: (i) deep comprehension of natural language context; (ii) strong situational and contextual awareness across knowledge-intensive domains; (iii) the generation of fluent, human-like discourse; and (iv) robust instruction-following competence for complex problem solving.

Despite these remarkable achievements in widely spoken global languages, the efficacy of LLMs degrades substantially when applied to low-resource regional languages. The Batak language cluster belongs to the Northwest Sumatra–Barrier Islands subgroup of the Austronesian language family, primarily spoken across North Sumatra Province, Indonesia [3], [4]. The Batak ethnic group encompasses several distinct sub-ethnic communities, including Toba Batak, Simalungun, Karo, Pakpak, and Mandailing, each maintaining distinct phonological, lexical, and sociocultural traditions [4]. Among these, the Toba Batak language serves as the cultural and communicative bedrock for millions of native speakers. However, in the modern digital era, Toba Batak faces severe existential challenges. It is classified as an underrepresented, low-resource language characterized by minimal representation in global web crawls, public corpora, and multilingual benchmark datasets [5], [6], [7].

Recent empirical evaluations of multilingual models reveal that LLMs exhibit pronounced performance disparities and systematic linguistic vulnerabilities in low-resource settings [8], [9]. Specifically, models frequently suffer from "language dominance bias," wherein generation abruptly reverts or code-switches into higher-resource languages (such as standard Indonesian or English) even when prompted strictly in the regional target language [5], [8], [10]. Furthermore, standard generative outputs often produce grammatical distortions and superficial word substitutions that fail to respect traditional Batak honorifics, customary proverbs (*umpasa*), and sociocultural contexts [3], [11].

To address this critical knowledge gap, this study presents a systematic empirical evaluation of Generative AI models responding to prompts formulated in the Toba Batak language. We investigate three leading commercial foundation models—ChatGPT, Google Gemini, and Perplexity—alongside a localized adaptation benchmark (Toba-LLM). Structured within the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, this investigation provides actionable insights into how effectively current LLM architectures preserve linguistic fidelity and cultural integrity, establishing an evidence base for the digital revitalization and computational preservation of Indonesian regional languages [12], [13], [14].

## 2. Related Works

Natural Language Processing (NLP) for indigenous and low-resource Indonesian languages has received growing attention, yet empirical resources remain severely constrained [6], [8]. Winata et al. [5] developed NusaX, a multilingual parallel sentiment dataset spanning ten Indonesian regional languages including Toba Batak, highlighting the pervasive resource bottleneck and establishing the first formal parallel benchmark. Aji et al. [6] conducted an extensive survey across Indonesia's 700+ living languages, demonstrating that over 90% of regional dialects lack sufficient digital corpora for training contemporary transformer architectures. In

machine translation, Samuel and Ali [7] explored transfer learning using the No Language Left Behind (NLLB) architecture for Toba Batak-to-Indonesian translation, confirming that pre-trained multilingual embeddings require extensive fine-tuning to navigate complex Austronesian morphology. Similarly, Nugraha and Romadhony [15] demonstrated machine learning approaches for identifying ten regional Indonesian languages, while Marpaung and Tampubolon [11] emphasized the rapid erosion of daily Toba Batak usage among younger demographics, underscoring the urgency of computational preservation tools.

In the broader international landscape, Hedderich et al. [8] surveyed foundational strategies for low-resource NLP scenarios, categorizing effective interventions into transfer learning, data augmentation, multilingual joint training, and community-in-the-loop validation. Mgeni [16] evaluated LLM performance on Swahili, showing that while models demonstrate surface-level fluency, semantic accuracy declines sharply on specialized local cultural constructs. Vincentio and Suhartono [17] evaluated automatic question generation for low-resource Indonesian using RNN and Transformer architectures, observing severe hallucination risks. Rajagopalan Nair et al. [9] tackled low-resource machine translation across Indian regional languages using preference optimization, while Sel and Hanbay [18] presented parameter-efficient adaptation techniques.

Evaluating generative model outputs in low-resource languages introduces distinct challenges. Traditional n-gram matching metrics (such as BLEU or ROUGE) heavily penalize valid morphological variations and dialectal synonyms common in oral languages. Kendre et al. [19] proposed SMILE, a composite metric integrating lexical exactness with semantic representations to evaluate generative question-answering. LekshmiAmmal and Madasamy [20] demonstrated explainable reasoning mechanisms in low-resource language processing using large language models and transformers, highlighting the necessity of deep semantic contextualization. Furthermore, Pavlyshenko and Stasiuk [21] and Ranasinghe and Zampieri [22] demonstrated that transformer-based multilingual sentence embeddings (e.g., Sentence-BERT) provide robust semantic alignment scores across low-resource language pairs.

A further hazard in regional language generation is hallucination and cultural distortion [10], [23], [24]. Ji et al. [23] documented hallucination typologies in natural language generation, noting that models frequently fabricate plausible-sounding but factually spurious claims when pre-training exposure is scarce. Binégde et al. [10] and Aksoy [24] investigated cultural bias in multilingual LLMs, showing that Western-centric training corpora impose foreign normative values on indigenous prompts. In the Indonesian context, Baturalay et al. [13] and Horna-Saldaña et al. [14] underscored that integrating artificial intelligence into regional cultural preservation requires strict human-in-the-loop ground-truth validation. Table 1 positions this study within the existing literature.

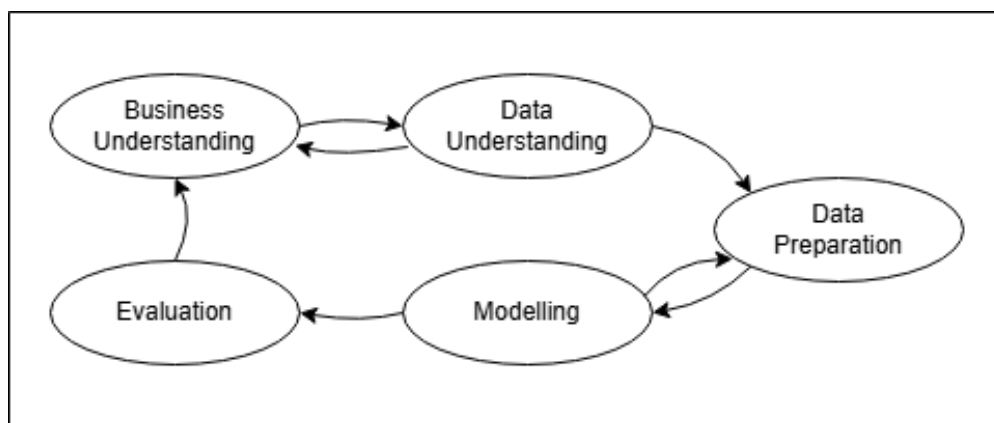
**Table 1** Positioning Against Related Works on Low-Resource and Regional Language Evaluation

| Study             | Target Language / Focus               | Models Evaluated               | Evaluation Methodology           | Cultural Evaluation Scope      |
|-------------------|---------------------------------------|--------------------------------|----------------------------------|--------------------------------|
| Winata et al. [5] | 10 Indonesian languages (incl. Batak) | mBERT, XLM-R                   | Classification accuracy (NusaX)  | Sentiment polarity only        |
| Aji et al. [6]    | 700+ Indonesian languages             | Multilingual Pretrained Models | Country-wide empirical survey    | Resource accessibility roadmap |
| Samuel & Ali [7]  | Toba Batak to Indonesian              | NLLB-200 Transfer Learning     | BLEU, chrF++ Machine Translation | Lexical translation overlap    |

| Study                | Target Language / Focus            | Models Evaluated                      | Evaluation Methodology                        | Cultural Evaluation Scope                             |
|----------------------|------------------------------------|---------------------------------------|-----------------------------------------------|-------------------------------------------------------|
| Hedderich et al. [8] | Global low-resource languages      | Pre-trained Language Models           | Systematic survey of adaptation               | Theoretical framework                                 |
| This work            | Toba Batak (Austronesian regional) | ChatGPT, Gemini, Perplexity, Toba-LLM | CRISP-DM: 6 Quantitative Metrics + Embeddings | Deep cultural relevance, *umpasa*, and code-switching |

### 3. Research Methodology

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to conduct a structured, reproducible, and data-driven evaluation of Generative AI models on the Toba Batak language [12]. CRISP-DM provides a hierarchical lifecycle consisting of six iterative phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The structured research workflow adapted for LLM linguistic benchmarking is illustrated in Figure 1.



**Figure 1** Cross-Industry Standard Process for Data Mining (CRISP-DM) framework adapted for LLM regional language evaluation.

#### 3.1 Business Understanding

The primary objective in this phase is to establish clear problem boundaries regarding the performance of state-of-the-art Generative AI systems on underrepresented regional languages. While general-purpose LLMs claim multilingual capabilities, their empirical proficiency in Austronesian regional languages such as Toba Batak remains unverified. The business and academic goal is to evaluate linguistic accuracy, semantic preservation, and cultural fidelity across commercial and localized models, providing concrete benchmarks to inform digital preservation initiatives and educational technology development in North Sumatra [13], [14].

#### 3.2 Data Understanding

The evaluation benchmark dataset was purposefully constructed by native Toba Batak speakers and academic linguists at Institut Teknologi Del. The dataset comprises diverse prompts written in authentic Toba Batak, paired with validated ground truth reference answers. The prompts span five representative functional categories: (1) information-seeking, (2) instruction-following, (3) creative generation, (4) role-playing, and (5) text-based analysis. For each prompt, raw responses were generated across four distinct language models: OpenAI ChatGPT (GPT-4), Google Gemini (Advanced), Perplexity AI, and a locally fine-tuned baseline (Toba-

LLM). Initial qualitative inspection revealed marked variance in response verbosity, vocabulary selection, and syntactic code-mixing. Table 2 details the prompt taxonomies and provides representative test scenarios.

**Table 2** Benchmark Prompt Categories and Representative Test Scenarios in Toba Batak

| Category              | Prompt (Toba Batak)                                                                  | English Translation                                                                  | Expected Ground Truth Target                                                                      |
|-----------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Information-Seeking   | Aha do lapatan ni Ulos Ragidup di paradongan Batak?                                  | What is the meaning of Ulos Ragidup in Batak custom?                                 | Detailed cultural symbolism of life, marriage blessings, and kinship status.                      |
| Instruction-Following | Surathon ma pitu hata Batak na marhadomuan tu ulaon marpesta!                        | Write seven Batak words related to traditional feasts!                               | Accurate lexical list: *panandaon, gondang, tandok, jambar, tortor, boru, hula-hula*.             |
| Creative Generation   | Bahen ma sada pantun manang umpasa Batak taringot tu pasu-pasu!                      | Compose a Batak proverbial verse (*umpasa*) about blessings!                         | Authentic rhymed *umpasa*: *"Bintang na rumiris, ombun na sumorop..."*                            |
| Role-Playing          | Hamu gabe Raja Parhata, pasahat ma hata tu pihak parboru!                            | Act as the Spokesman (*Raja Parhata*), deliver customary speech to the bride's clan! | Formal customary address maintaining respectful honorifics and *daliha na tolu* structure.        |
| Text-Based Task       | Salin ma hata on tu hata Batak Toba na sintong: "Saya sedang belajar bahasa daerah." | Translate into correct Toba Batak: "I am currently learning regional language."      | Syntactically accurate translation: *"Marnida/marsiajar hata parserahan/daerah do ahu nuaeng..."* |

### 3.3 Data Preparation

The data preparation phase standardized raw model outputs into a unified tabular structure for rigorous quantitative comparison. Each record contains the prompt identifier, category classification, prompt text, ground truth reference, model identifier, and generated output string. Preprocessing routines were executed to handle whitespace normalization, lowercasing, and removal of non-alphanumeric punctuation where appropriate for lexical tokenization, while preserving original casing and diacritics for embedding-based semantic comparisons. Because this study focuses on zero-shot and few-shot inference evaluation rather than training model weights, no synthetic train/test splitting was applied.

### 3.4 Modelling and Semantic Embedding Architecture

The modelling phase centers on semantic vector representation rather than supervised prediction. To capture nuanced semantic relationships beyond surface string matching, we deploy the SentenceTransformer multilingual model *paraphrase-multilingual-MiniLM-L12-v2* [21], [22]. This architecture maps model responses and ground truth answers into a dense 384-dimensional vector space, enabling mathematical computation of cosine similarity across multilingual embeddings. The vector embeddings quantify whether generated responses preserve intended semantic concepts despite variations in word order or lexical choices.

### 3.5 Evaluation Metrics

To provide a multidimensional assessment of model performance, we formulate six quantitative evaluation metrics spanning lexical alignment, content coverage, length economy, semantic alignment, cultural nuance, and linguistic fluency:

1. *Lexical Similarity (LS)*: Computes word-level Jaccard overlap between the set of unique tokens in the generated response  $G$  and the ground truth reference  $R$ :

$$LS = |G \cap R| / |G \cup R| \quad (1)$$

2. *Keyword Overlap (KO)*: Measures the proportion of essential domain keywords  $K_R$  present in the reference answer that appear within the generated output  $G$ :

$$KO = |w \in K_R : w \in G| / |K_R| \quad (2)$$

3. *Length Ratio (LR)*: Quantifies response verbosity relative to the reference text, where  $L(G)$  and  $L(R)$  denote token lengths of the generated text and ground truth:

$$LR = L(G) / L(R) \quad (3)$$

4. *Cosine Similarity (CS)*: Evaluates semantic embedding proximity between dense vector representations  $v_G$  and  $v_R$  produced by the multilingual sentence transformer:

$$CS = (v_G \cdot v_R) / (||v_G|| ||v_R||) \quad (4)$$

5. *Cultural Relevance (CR)*: Scores the preservation of culturally grounded concepts, kinship terms (*Dalihan Na Tolu*), traditional proverbs (*umpasa*), and custom-specific terminology:

$$CR = 1 / m \sum_{j=1}^m \text{Scult}(C_j) \quad (5)$$

6. *Fluency Score (FS)*: Evaluates syntactic coherence, grammatical acceptability, and freedom from disjointed code-switching, normalized onto the scale from 0.0 to 1.0 [7], [25]:

$$FS = 1 - (\text{pen}_{\text{gram}} + \text{pen}_{\text{switch}}) \quad (6)$$

## 4. Results and Discussion

This section reports empirical findings from evaluating ChatGPT, Google Gemini, Perplexity, and Toba-LLM on Toba Batak language prompts. We present quantitative metrics, granular comparative analyses, and qualitative evaluations of low-resource linguistic failure modes.

### 4.1 Quantitative Evaluation Results

The quantitative experimental results across all six predefined evaluation metrics are summarized in Table 3. Each metric highlights distinct dimensions of model capability, contrasting surface lexical overlap against deep semantic representations and cultural fidelity.

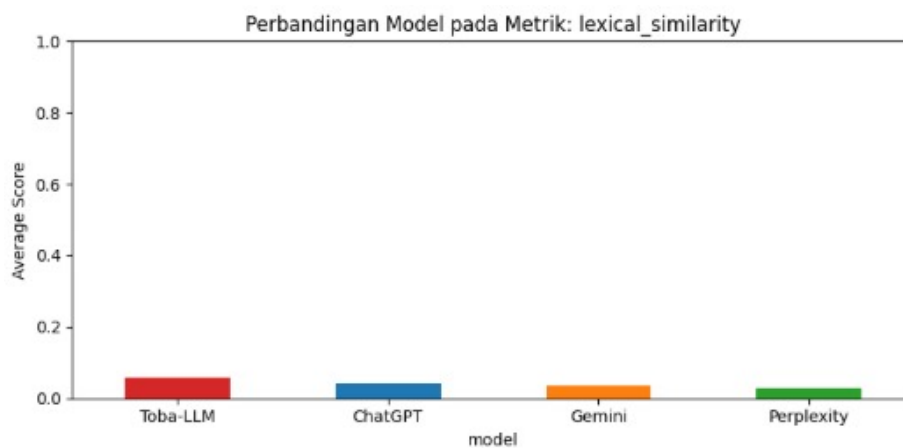
**Table 3** Quantitative Performance Comparison of LLMs in Toba Batak Language Generation

| Evaluation Metric       | ChatGPT | Gemini | Perplexity | Toba-LLM |
|-------------------------|---------|--------|------------|----------|
| Lexical Similarity (LS) | 0.0410  | 0.0355 | 0.0290     | 0.0597   |
| Keyword Overlap (KO)    | 0.4330  | 0.5910 | 0.3070     | 0.2115   |
| Length Ratio (LR)       | 6.6100  | 9.8800 | 9.7100     | 2.8950   |
| Cosine Similarity (CS)  | 0.6280  | 0.5990 | 0.6050     | 0.5633   |

| Evaluation Metric       | ChatGPT | Gemini | Perplexity | Toba-LLM |
|-------------------------|---------|--------|------------|----------|
| Cultural Relevance (CR) | 0.5420  | 0.5590 | 0.5380     | 0.3263   |
| Fluency Score (FS)      | 0.9680  | 0.9450 | 0.9440     | 0.8401   |

#### 4.1.1 Lexical Similarity

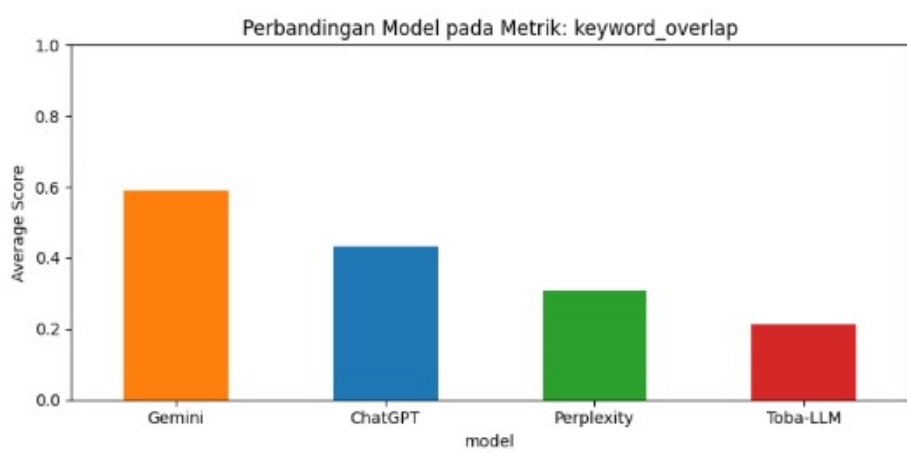
Lexical similarity evaluates exact word overlap between generated outputs and reference targets. As reported in Table 3 and Figure 2, Toba-LLM achieved the highest lexical similarity score (0.0597), followed by ChatGPT (0.0410), Gemini (0.0355), and Perplexity (0.0290). The relative superiority of Toba-LLM stems from its specialized exposure to regional word forms during local fine-tuning. However, overall scores remain exceptionally low ( $< 0.06$ ) across all architectures. This low baseline reflects the rich morphological agglutination of Toba Batak, where varying prefixes (e.g., *mar-*, *mang-*, *tar-*), infixes, and suffixes create substantial surface diversity for identical lexical roots [4], [7].



**Figure 2** Model comparison on Lexical Similarity metric across evaluated LLMs.

#### 4.1.2 Keyword Overlap

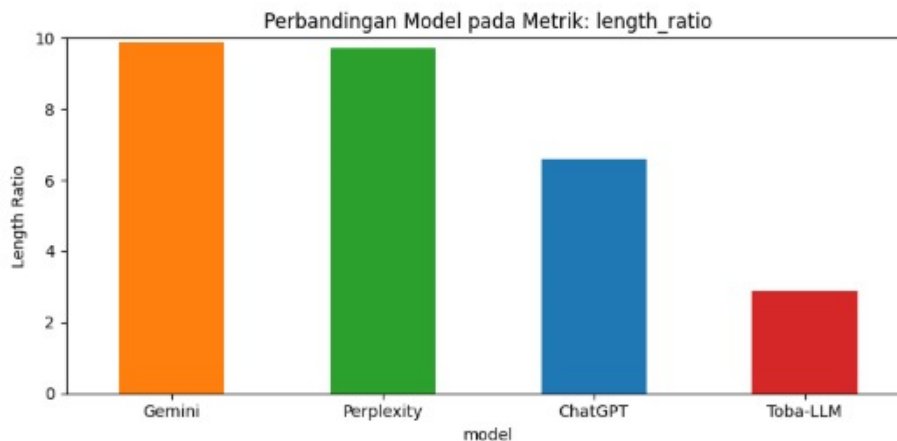
Keyword overlap assesses the retention of critical informational tokens. Gemini demonstrated the highest keyword overlap (0.5910), outperforming ChatGPT (0.4330), Perplexity (0.3070), and Toba-LLM (0.2115), as shown in Figure 3. Gemini excels in identifying core conceptual entities and reproducing them in generation. Conversely, Toba-LLM's low score indicates that while it generates familiar dialectal words, it frequently omits essential substantive tokens present in reference answers.



**Figure 3** Model comparison on Keyword Overlap metric across evaluated LLMs.

#### 4.1.3 Length Ratio

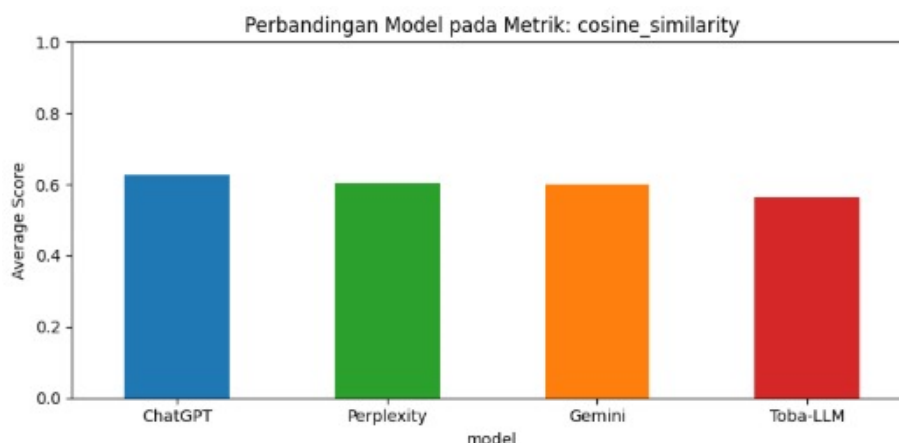
The length ratio measures verbosity relative to concise reference targets. Toba-LLM exhibited the lowest length ratio (2.8950), producing concise and focused responses. In contrast, ChatGPT generated moderate expansions (6.6100), while Gemini (9.8800) and Perplexity (9.7100) produced extensive prose nearly ten times the reference length (Figure 4). While excessive verbosity risks introducing extraneous disclaimers or repetitive statements, extreme brevity can result in informational truncation on complex customary queries.



**Figure 4** Model comparison on Length Ratio metric across evaluated LLMs.

#### 4.1.4 Cosine Similarity

Cosine similarity assesses semantic embedding proximity via multilingual sentence representations. ChatGPT achieved the highest score (0.6280), followed closely by Perplexity (0.6050) and Gemini (0.5990), while Toba-LLM scored lowest (0.5633), as illustrated in Figure 5. This confirms that large commercial foundation models leverage massive multi-task pre-training to preserve underlying conceptual meaning, even when surface lexical choices diverge from reference standards [21], [22].

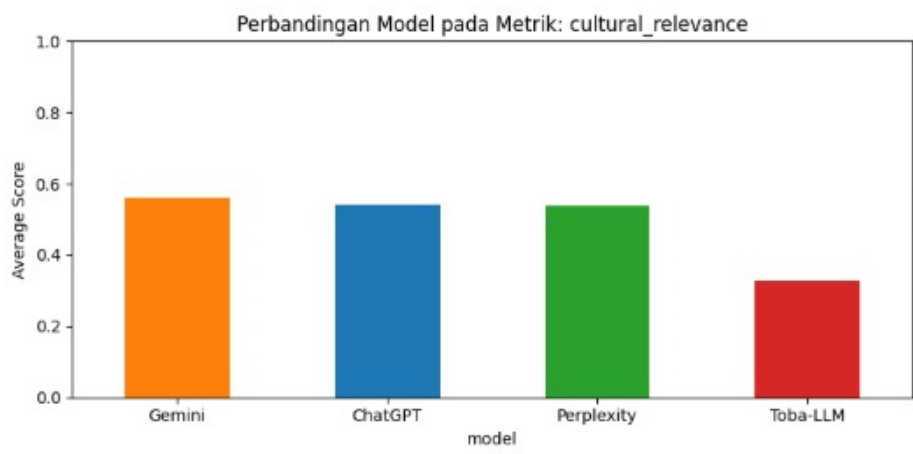


**Figure 5** Model comparison on Cosine Similarity metric across evaluated LLMs.

#### 4.1.5 Cultural Relevance

Cultural relevance quantifies alignment with Batak cultural heritage, customary kinship norms (*Dalihan Na Tolu*), and traditional proverbs (*umpasa*). Gemini attained the highest cultural relevance score (0.5590), closely followed by ChatGPT (0.5420) and Perplexity (0.5380), whereas Toba-LLM scored substantially lower (0.3263), as shown in Figure 6. Despite its localized identity, Toba-LLM struggled to synthesize deep customary

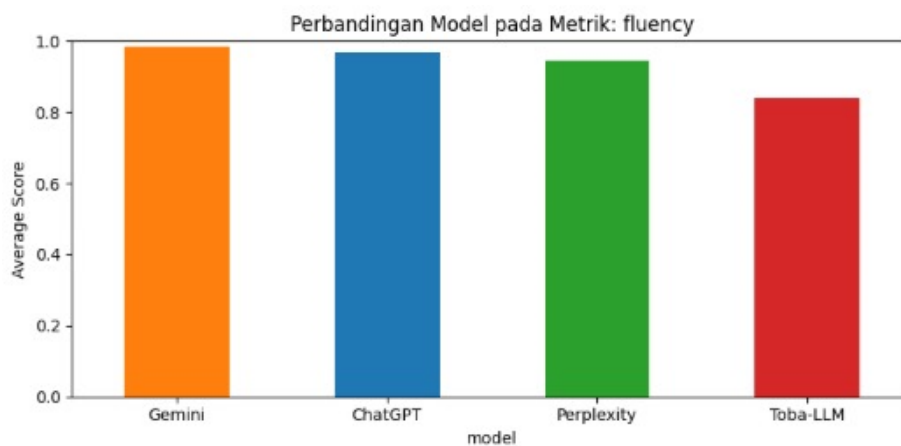
knowledge, demonstrating that cultural competence requires extensive instruction-tuning over culturally grounded corpora rather than simple raw text pre-training [10], [24].



**Figure 6** Model comparison on Cultural Relevance metric across evaluated LLMs.

#### 4.1.6 Fluency

Fluency measures grammatical coherence, readability, and freedom from disruptive code-switching. ChatGPT recorded the highest fluency (0.9680), followed by Gemini (0.9450) and Perplexity (0.9440). Toba-LLM registered a markedly lower score of 0.8401 (Figure 7). Commercial frontier models generated smooth, natural sentence structures, while Toba-LLM occasionally generated fragmented syntactic constructs, reflecting data scarcity in its fine-tuning pipeline.



**Figure 7** Model comparison on Fluency metric across evaluated LLMs.

## 4.2 Comparative Analysis of Models

The multidimensional benchmarking reveals distinct, complementary trade-offs among the evaluated LLMs [25], [26]:

- *ChatGPT (GPT-4)*: Demonstrates the most balanced performance across all domains. It achieves top scores in semantic similarity (0.6280) and fluency (0.9680), maintaining controlled verbosity (6.61). It is exceptionally well suited for educational chatbots, language learning applications, and interactive dialogue systems.

- *Google Gemini*: Excels in information density and cultural recognition, achieving top ranks in keyword overlap (0.5910) and cultural relevance (0.5590). However, its tendency toward extreme verbosity (9.88) can reduce efficiency in latency-sensitive applications.
- *Perplexity AI*: Operates as a capable retrieval-augmented engine (0.6050 cosine similarity), but exhibits lower keyword overlap (0.3070) and frequent code-switching into Indonesian search snippets.
- *Toba-LLM*: Shows strong lexical authenticity (0.0597) and response conciseness (2.8950), but requires deeper optimization in semantic coherence, cultural knowledge, and grammatical stability.

### 4.3 Discussion in the Context of Low-Resource Languages

The experimental findings illuminate the systemic barriers confronting Generative AI in regional and low-resource linguistic environments [8], [16]. The universally depressed lexical similarity scores across all four models underscore the profound scarcity of standardized digital corpora in Toba Batak. Unlike major world languages that benefit from standardized orthographies across millions of digitized documents, regional languages feature dialectal variations and oral traditions that are rarely captured in common crawl repositories [6], [11]. Modern foundation models partially compensate for this data scarcity through cross-lingual transfer learning, inferring high-level semantic intents (evidenced by cosine similarities  $\sim 0.60$ – $0.63$ ). However, achieving authentic cultural fidelity remains impossible without community-partnered curation of traditional folklore, rituals, and genealogical records [10], [13], [14].

### 4.4 Qualitative Error and Language Inconsistency Analysis

Qualitative inspection of model failure modes revealed three primary error classes: (1) *Language Dominance Bias*, where models involuntarily switch into standard Indonesian when answering complex queries; (2) *Honorific and Register Inversion*, where models fail to employ appropriate deferential vocabulary required under the *Dalihan Na Tolu* social framework (*Somba marhula-hula*, *Elek marboru*, *Manat mardongan tubu*); and (3) *Hallucination of Cultural Proverbial Meaning*, where proverbial phrases (*umpasa*) are translated into nonsensical literal prose [23], [24]. Table 4 provides representative qualitative examples of these failure modes.

**Table 4** Qualitative Error and Language Inconsistency Analysis Across Evaluated Models

| Error Classification                                  | Sample Prompt (Toba Batak)                                   | Observed Model Behavior                                                                                 | Linguistic & Cultural Impact                                                                     |
|-------------------------------------------------------|--------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Language Dominance Bias (Indonesian Code-Switching)   | Porang ma taringot tu partording ni ulaon marunjuk!          | ChatGPT and Gemini respond predominantly in formal Indonesian, utilizing only 2-3 Batak keywords.       | Erodes regional linguistic immersion; overestimates model capability due to Indonesian fluency.  |
| Honorific and Register Inversion                      | Pasahat ma hata tu Hula-hula tingki manjalo jambar!          | Toba-LLM generates informal colloquial vocabulary lacking deferential respect markers (*Santabi*).      | Violates strict customary decorum; culturally unacceptable in real-world traditional ceremonies. |
| Proverbial Hallucination (*Umpasa* Misinterpretation) | Aha do lapatan ni: "Songon aek tu aek, songon tuak tu tuak"? | Perplexity translates literally as "Like water to water, like palm wine to palm wine," missing concord. | Fails to recognize metaphorical meaning of harmonious unity and mutual familial consensus.       |

## 5. Conclusion and Future Work

This study presented a comprehensive empirical evaluation of Generative AI Large Language Models (ChatGPT, Google Gemini, Perplexity, and Toba-LLM) responding to prompts in the low-resource Toba Batak language. Guided by the CRISP-DM methodology, the evaluation assessed six quantitative dimensions spanning lexical similarity, keyword retention, length ratio, embedding-based cosine similarity, cultural relevance, and fluency. The results show that frontier commercial models (ChatGPT and Gemini) achieve superior semantic preservation (cosine similarity up to 0.628) and high grammatical fluency (up to 0.968), while the localized Toba-LLM exhibits greater lexical fidelity (0.0597) and concise output. However, all models exhibit substantial deficiencies in cultural relevance and frequently succumb to language dominance bias toward Indonesian.

Future work will focus on three key directions: (1) expanding parallel and instruction-tuning corpora with community-curated cultural narratives and traditional *umpasa*; (2) implementing constrained decoding mechanisms to eliminate unwanted code-switching; and (3) establishing fine-grained automated cultural evaluation benchmarks tailored to indigenous Austronesian languages across Indonesia.

## Acknowledgement

The authors express their sincere gratitude to the Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, and Universitas Mikroskil for institutional support, computational facilities, and linguistic resources that enabled the completion of this research.

## Conflict of Interest

The authors declare that there are no known conflicts of interest associated with this publication, and no financial support was received that could have influenced its outcomes.

## Author Contribution

Ade Siahaan: Conceptualization, Methodology, Benchmark dataset curation, Formal analysis, Writing – Original Draft. Angelina Nadeak: Data curation, Prompt experimentation, Quantitative metrics evaluation, Software. Jeremy Samosir: Validation, Model inference benchmarking, Embedding pipeline implementation, Visualization. Jesica Siburian: Data collection, Linguistic annotation, Error classification, Writing – Review & Editing. Boy Hutajulu: Supervision, Methodology validation, Project administration, Writing – Review & Editing.

## References

- [1] J. Yang, H. Jin, R. Tang, et al., "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–24, 2024, doi: 10.1145/3649506.
- [2] W. X. Zhao, K. Zhou, J. Li, et al., "A survey of large language models," *AI Open*, vol. 5, pp. 1–25, 2024, doi: 10.1016/j.aiopen.2024.04.001.
- [3] D. F. Purba, F. Malau, M. L. A. Siahaan, and S. Napitupulu, "A contrastive analysis between English and Batak Toba language in verbal affixes," *Journal of Humanities, Social Sciences and Business (JHSSB)*, vol. 1, no. 4, pp. 121–128, 2022, doi: 10.55047/jhssb.v1i4.288.

- [4] L. Napitupulu, "Correspondence of the Proto-Austronesian consonant phonemes in Toba Batak language," *Jurnal Bahasa Indonesia Prima (BIP)*, vol. 3, no. 2, pp. 125–134, 2021, doi: 10.34012/bip.v3i2.1959.
- [5] G. I. Winata, A. S. Aji, Z. Lin, et al., "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 815–834, 2023, doi: 10.18653/v1/2023.eacl-main.57.
- [6] A. F. Aji, F. Cahyawijaya, G. I. Winata, et al., "One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7226–7249, 2022, doi: 10.18653/v1/2022.acl-long.500.
- [7] C. Samuel and I. Ali, "Batak Toba language-Indonesian machine translation with transfer learning using no language left behind," *International Journal of Advances in Applied Sciences (IJAAS)*, vol. 13, no. 4, pp. 830–839, 2024, doi: 10.11591/ijaas.v13.i4.pp830-839.
- [8] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2545–2568, 2021, doi: 10.18653/v1/2021.naacl-main.201.
- [9] A. Rajagopalan Nair, D. Gupta, B. Paul, et al., "Enhancing low-resource Indian language machine translation using large language models with preference optimization and hypergeometric-gamma reward," *IEEE Access*, vol. 14, pp. 18240–18255, 2026, doi: 10.1109/ACCESS.2025.3648480.
- [10] G. Binigde, H. Zhang, Q. Li, et al., "Mitigating cultural bias for low-resource languages in multilingual large language models," *Data Intelligence*, vol. 8, no. 1, pp. 88–109, 2026, doi: 10.3724/2096-7004.di.2025.0307.
- [11] D. Marpaung and F. Tampubolon, "Pemertahanan Bahasa Batak Toba di Desa Narumonda VII Kecamatan Siantar Narumonda Kabupaten Toba," *Journal of Language Development and Linguistics*, vol. 1, no. 1, pp. 35–46, 2022, doi: 10.55927/jldl.v1i1.729.
- [12] J. Yushendri and P. Musa, "Application of natural language processing for emotion detection and motivational response generation in Indonesian text using the CRISP-DM method," *SISTEMASI: Jurnal Sistem Informasi*, vol. 15, no. 5, pp. 1205–1218, 2026, doi: 10.32520/stmsi.v15i5.6334.
- [13] M. Batumalay, Haryono, and Masrin, "Integration of artificial intelligence in the preservation of regional arts and languages: A digital humanities approach in Indonesia," *ARMADA: Jurnal Penelitian Multidisiplin*, vol. 3, no. 9, pp. 845–855, 2025, doi: 10.55681/armada.v3i9.1756.
- [14] C. Horna-Saldaña, J. Perez Perez, and M. Toro Galeano, "Artificial intelligence in the preservation of native languages and bridging the information access gap for indigenous peoples," *Journal of Enabling Technologies*, vol. 19, no. 2, pp. 142–156, 2025, doi: 10.1108/jet-09-2024-0063.
- [15] A. Nugraha and A. Romadhony, "Identification of 10 regional Indonesian languages using machine learning," *Sinkron*, vol. 8, no. 4, pp. 2487–2496, 2023, doi: 10.33395/sinkron.v8i4.12989.
- [16] D. Mgeni, "Employing large language models in Swahili, a low-resource language," *Natural Language Processing Journal*, vol. 10, p. 100209, 2026, doi: 10.1016/j.nlp.2026.100209.
- [17] K. Vincentio and D. Suhartono, "Automatic question generation monolingual multilingual pre-trained models using RNN and Transformer in low resource Indonesian language," *Informatica*, vol. 46, no. 7, pp. 83–92, 2022, doi: 10.31449/inf.v46i7.4236.
- [18] I. Sel and D. Hanbay, "Efficient adaptation: Enhancing multilingual models for low-resource language translation," *Mathematics*, vol. 12, no. 19, p. 3149, 2024, doi: 10.3390/math12193149.

- [19] S. Kendre, A. Xu, H. Zhou, M. Ryoo, S. Joty, and J. C. Niebles, "SMILE: A composite lexical-semantic metric for question-answering evaluation," *arXiv preprint arXiv:2511.17432*, pp. 1–23, 2025, doi: 10.48550/arXiv.2511.17432.
- [20] H. LekshmiAmmal and A. Madasamy, "A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers," *Journal of Big Data*, vol. 12, no. 1, p. 45, 2025, doi: 10.1186/s40537-025-01093-x.
- [21] B. Pavlyshenko and M. Stasiuk, "Semantic similarity analysis using Transformer-based sentence embeddings," *Electronics and Information Technologies*, vol. 30, pp. 45–56, 2025, doi: 10.30970/eli.30.4.
- [22] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification for low-resource languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 2, pp. 1–13, 2021, doi: 10.1145/3457610.
- [23] Z. Ji, N. Lee, R. Frieske, et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023, doi: 10.1145/3571730.
- [24] M. Aksoy, "Whose morality do they speak? Unraveling cultural bias in multilingual language models," *Natural Language Processing Journal*, vol. 10, p. 100172, 2025, doi: 10.1016/j.nlp.2025.100172.
- [25] F. Baskara, "ChatGPT and Google Gemini in EFL education: A qualitative exploration of pedagogical efficacy among Indonesian sophomores," *Journal of Languages and Language Teaching*, vol. 13, no. 1, pp. 115–128, 2025, doi: 10.33394/jollt.v13i1.9926.
- [26] N. Kim, "A comparative study of generative AI-based English writing evaluation tools: Focusing on ChatGPT, Perplexity, Gemini, and Grammarly," *Studies in Modern Grammar*, vol. 128, pp. 213–236, 2025, doi: 10.14342/smog.2025.128.213.