When Words Don't Mean What They Say: Figurative Understanding in Bengali Idioms

Adib Sakhawat[1] **Shamim Ara Parveen**[2] **Md Ruhul Amin**[2]
Shamim Al Mahmud[2] **Md Saiful Islam**[2] **Tahera Khatun**[2]

[1]Islamic University of Technology [2]National University of Bangladesh
adibsakhawat@iut-dhaka.edu

Abstract

Figurative language understanding remains a significant challenge for Large Language Models (LLMs), especially for low-resource languages. To address this, we introduce the *Bangla Bagdhara* dataset, a large-scale, culturally-grounded corpus of 10,361 Bengali idioms. Each idiom is annotated under a comprehensive 19-field schema, established and refined through a deliberative expert consensus process that captures its semantic, syntactic, cultural, and religious dimensions, providing a rich, structured resource for computational linguistics. To establish a robust benchmark for Bangla figurative language understanding, we evaluate 30 state-of-the-art multilingual and instruction-tuned LLMs on the task of inferring figurative meaning. Our results reveal a critical performance gap, with no model surpassing 50% accuracy, a stark contrast to significantly higher human performance (83.4%). This underscores the limitations of existing models in cross-linguistic and cultural reasoning. By releasing the *Bangla Bagdhara* dataset and benchmark, we provide foundational infrastructure for advancing figurative language understanding and cultural grounding in LLMs for Bengali and other low-resource languages.

Keywords: Bengali Idioms, Figurative Language Understanding, Linguistic Dataset, Cultural Annotation, Low-Resource NLP, LLM Evaluation

1. Introduction

The computational modeling of figurative language presents a significant and persistent challenge in Natural Language Processing (NLP). As noted in recent surveys, research on idioms has proceeded along two parallel tracks, psycholinguistics and computational linguistics, with little crossover (Flor et al., 2025). While high-resource languages like English have benefited from extensive corpus development for both type-based idiom detection (e.g., MAGPIE (Haagsma et al., 2020)) and token-based contextual disambiguation (e.g., IDIX (Sporleder et al., 2010)), this progress remains starkly uneven. The difficulty of idiom annotation, even in high-resource settings (Street et al., 2010), is compounded for the over 1.5 billion speakers of South Asian languages, which remain critically under-resourced in NLP (Arora et al., 2022).

Bangla (Bengali), the world's seventh most spoken language, epitomizes this resource scarcity. Prior work on Bangla figurative language has been limited in scope and focus. Early efforts centered on rule-based translation systems for small sets of idioms (Khatun et al., 2020) or statistical identification of general multi-word expressions (MWEs) (Chakraborty et al., 2014). More recently, datasets for figurative language have emerged, such as Alankaar (Rakshit and Flanigan, 2025), which provides a valuable resource of 2,500 examples for metaphor identification. However, a large-scale, computationally-oriented *idiom* corpus with deep cultural annotation has been unavailable. Existing resources are either smaller in scale, focus on different linguistic phenomena like metaphors, or lack the cultural grounding necessary to interpret idioms in context, a difficulty magnified by Bangla's complex morphology (Deo and Sharma, 2006).

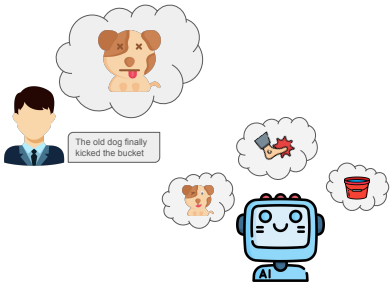

Figure 1: Conceptual illustration contrasting human and AI interpretations of idioms ('kicked the bucket').

While existing multilingual resources offer some utility, they possess notable limitations. Datasets like LIdioms (Moussallem et al., 2018) and IMIL (Agrawal et al., 2018) focus primarily on facilitating cross-lingual transfer rather than providing the monolingual depth and fine-grained cultural annotation necessary for a nuanced understanding of idioms in their native context. Similarly, commend-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6870–6879
11–16 May 2026.

able regional efforts such as the Konidioms corpus for Konkani (Shaikh et al., 2024) operate at a scale insufficient for developing and evaluating robust computational models.

To address this critical gap, we introduce the *Bangla Bagdhara* dataset, a large-scale, culturally-grounded corpus of 10,361 Bangla idioms. This work offers four principal contributions:

1. **Scale and Coverage:** With 10,361 meticulously annotated entries, the *Bangla Bagdhara* dataset is the largest and most comprehensive figurative language resource for Bangla.

2. **Rich Annotation Schema:** We introduce a novel 19-field annotation framework. It includes literal and figurative meanings (Bangla/English), usage examples, and deep cultural metadata (historical origins, religious context, socio-spatial relevance).

3. **Extensive Benchmarking:** We establish the first comprehensive benchmark for Bangla figurative language understanding. We evaluated 30 state-of-the-art LLMs on a subset of **100 idioms hand-picked by expert consensus**. For a human baseline, **15 Bangladeshi high school (6) and undergrad (9) students** took the test in a **controlled environment**, achieving a **mean score of 83.4%**. The results show a stark performance chasm: while humans averaged 83.4% accuracy, no LLM surpassed the 50% mark, exposing a critical deficiency in current models.

4. **Cultural Grounding:** The *Bangla Bagdhara* dataset systematically integrates cultural context as a primary annotation layer, enabling novel research at the intersection of computational linguistics and cultural analytics.

The *Bangla Bagdhara* dataset and our accompanying benchmark analysis provide foundational infrastructure to advance a new generation of culturally-aware NLP applications for Bangla, including machine translation, sentiment analysis, and educational tools. By making this resource publicly available, we aim to catalyze targeted research efforts to bridge the significant performance gap in multilingual figurative language understanding.[1]

2. Related Work

Computational modeling of idiomaticity presents significant challenges, particularly for resource-constrained languages with substantial linguistic diversity. While English-centric approaches dominate the field, recent multilingual expansions remain inadequate for languages such as Bangla. A survey of existing literature reveals critical gaps in scale, annotation depth, and cultural metadata that our work aims to address (Flor et al., 2025).

The **MAGPIE corpus** (Haagsma et al., 2020) established a benchmark with over 50,000 English expressions, demonstrating the feasibility of large-scale idiom annotation. The **FLUTE dataset** (Chakrabarty et al., 2022) introduced a paradigm shift through textual explanations across 9,000 figurative instances using human-AI collaboration. Similarly, the **IDIX corpus** provided context-aware annotations for English idioms, which facilitated token-level disambiguation of literal versus figurative usage (Sporleder et al., 2010). Work by (Street et al., 2010) on the American National Corpus further highlighted the subjective nature of idiom identification and the importance of including varied syntactic structures, such as prepositional phrases.

Multilingual efforts have achieved limited success. The **LIdioms dataset** (Moussallem et al., 2018) provided cross-lingual linkages for 815 idioms across five languages, while the **IMIL corpus** (Agrawal et al., 2018) expanded to 2,200 idioms across eight Indian languages. However, these resources remain limited in both scale and annotation depth compared to their English counterparts.

Corpus	Size	Languages	Annotation Fields
MAGPIE	50,000+	1 (English)	5
FLUTE	9,000	1 (English)	8
LIdioms	815	5	6
IMIL	2,200	8	7
Konidioms	1,597	1 (Konkani)	9
Alankaar	2,500	1 (Bangla)	2
Bangla Bagdhara	**10,361**	**1 (Bangla)**	**19**

Table 1: Comparative Analysis of Major Idiom Corpora

Significant regional disparities persist, with South Asian languages severely underrepresented despite their substantial speaker populations. While Konkani received attention through the **Konidioms corpus** (Shaikh et al., 2024) with 1,597 entries, Bengali serving over 300 million speakers has lacked dedicated computational resources. Early work on Bengali multi-word expressions (MWEs) by (Chakraborty et al., 2014) highlighted their non-compositional nature, and a rule-based translation system for Bengali idioms was proposed by (Khatun et al., 2020). More recently, the **Alankaar** dataset provided 2,500 annotated examples of metaphors in Bangla, exploring the cultural challenges of translating figurative language (Rakshit and Flanigan, 2025). However, a large-scale, comprehensively annotated idiom-specific resource has been missing. This scarcity reflects broader challenges in South Asian NLP (Baker et al., 2003), where

[1] https://www.kaggle.com/datasets/sakhadib/bangla-bagdhara

Bangla's linguistic complexity (Deo and Sharma, 2006) and semantic non-compositionality intersect with cultural contexts (Arora et al., 2022).

Annotation methodologies have evolved from basic frameworks like **IDIX** (Sporleder et al., 2010) and **PIE** (Haagsma et al., 2020) to sophisticated multimodal approaches with **V-FLUTE** (Chakrabarty et al., 2022). However, contemporary datasets often focus on narrow semantic distinctions, with FLUTE's explanations constrained to natural language inference tasks and LIdioms prioritizing cross-linguistic connectivity over semantic depth.

Cultural metadata remains critically underdeveloped. While some datasets include basic domain classifications, systematic annotation of historical significance, religious context, and socio-spatial dimensions is largely absent. Cross-linguistic approaches frequently prioritize direct translation over cultural adaptation (Flor et al., 2025), neglecting essential cultural models. Recent research emphasizes culturally informed annotation schemas (Pakray et al., 2025) for morphologically complex languages.

Both open-source and proprietary large language models significantly struggle with understanding figurative language (Khoshtab et al., 2025), particularly for low-resource languages due to training data gaps. Open-source models demonstrate particularly pronounced limitations compared to some proprietary systems (Khoshtab et al., 2025).

The development of *Bangla Bagdhara* addresses these gaps for Bangla through a large-scale, comprehensively annotated idiom resource with a 19-field schema, establishing new standards for low-resource language resource development.

3. Methodology

The construction of the *Bangla Bagdhara* dataset and its subsequent evaluation followed a systematic, multi-phase protocol designed to ensure linguistic authenticity, computational robustness, and scholarly integrity. Figure 2 shows the complete pipeline of the *Bangla Bagdhara* dataset construction, annotation, and evaluation process.

3.1. Dataset Architecture and Schema

The dataset's core is a comprehensive 19-field JSON schema, engineered to capture the multidimensional nature of idiomatic expressions beyond semantics. Each of the 10,361 entries adheres to this structure, which is organized into four key logical groups:

1. **Semantic and Cross-Lingual Mapping:** Fields such as `idiom`, `literal_meaning`, `figurative_meaning_bn/en`, and `similar_in_english` establish a triad mapping denotation to connotation, providing cross-lingual anchors for translation and analogical reasoning.
2. **Pragmatic Contextualization:** `example_sentences_in_bangla/english`, `usage_domain`, and `tags` situate idioms in authentic linguistic environments, which is essential for training context-aware models.
3. **Sociocultural and Affective Metadata:** Boolean flags for `historical_significance`, `religious_significance`, and `cultural_significance`, alongside `sentiment` polarity, encode vital symbolic and affective context.
4. **Diachronic and Ethnolinguistic Information:** The `scape` field defines the socio-spatial domain (e.g., urban, rural), while `history` and `note` preserve provenance and expert ethnolinguistic commentary.

The complete schema is detailed in Table 2.

Field	Type	Language
id	integer	English
idiom	string	Bangla
alternative_idioms	array	Bangla
literal_meaning	string	Bangla
figurative_meaning_bn	string	Bangla
figurative_meaning_en	string	English
similar_in_english	array	English
example_sentences_in_bangla	array	Bangla
example_sentences_in_english	array	English
usage_domain	array	Mixed
tags	array	Mixed
frequency	string	English
sentiment	string	English
historical_significance	boolean	English
religious_significance	boolean	English
cultural_significance	boolean	English
scape	string	English
history	array	Mixed
note	string	Mixed

Table 2: The 19-field schema for each entry in the *Bangla Bagdhara* dataset.

3.2. Corpus Construction and Annotation Protocol

The corpus was developed over five months by a dedicated, multidisciplinary team of six researchers with specific domain expertise.

Phase 1: Data Collection and Filtering. We aggregated an initial corpus of idioms from authoritative sources spanning three centuries, including

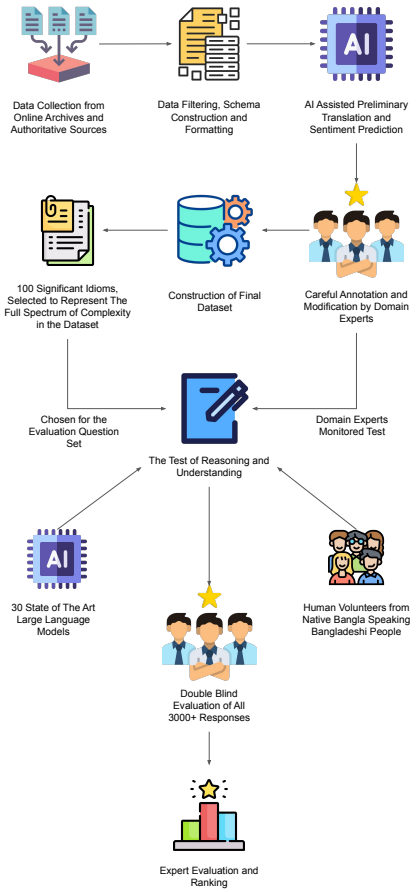

Figure 2: The complete pipeline of the *Bangla Bagdhara* dataset construction, annotation, and evaluation process.

Bangla Academy dictionaries, regional folk literature, and digital archives. The expert team then manually filtered this collection to verify authenticity and relevance, resulting in 10,361 entries that were subsequently normalized to Standard Bangla Unicode (UTF-8).

Phase 2: AI-Assisted Annotation with Human Oversight. We employed a human-in-the-loop workflow where AI served strictly as an assistive tool. For each entry, models like Gemini-2.5-flash and GPT-4o generated initial draft translations and sentiment suggestions. Crucially, this AI-generated content was never accepted directly. Instead, it served as a baseline for a meticulous human review process where at least two domain experts from the team would review, validate, and extensively refine or entirely replace the suggestions to ensure linguistic precision and cultural nuance.

Phase 3: Expert Validation and Scholarly Consensus. The final phase prioritized deep qualitative accuracy through direct engagement with a specialized expert panel, deliberately eschewing standard crowd-sourcing protocols. This panel was composed of six members, bringing a blend of academic theory and pedagogical practice: one computer scientist and five experts in Bangla language and linguistics (including academics and educators).

Validation by this group was not conducted via individual scoring rubrics; instead, consensus was actively constructed through iterative physical meetings and dedicated online discussions. This methodology was chosen to leverage the deep, non-technical expertise of the scholars, allowing for the resolution of ambiguities in a way that quantitative metrics cannot capture. Every single data point in the corpus, regardless of its origin, was subject to this review. For an idiom entry to be finalized and marked as "done", a minimum of two experts had to reach complete agreement through these discussions. This process ensures that the final dataset reflects deep scholarly consensus rather than a statistical aggregation of disconnected annotations.

3.3. Inter-Annotator Agreement and Validation Protocol

The annotation process followed a qualitative consensus-based validation framework rather than independent parallel annotation with post-hoc statistical aggregation. For each idiom entry, annotations were reviewed by a minimum of two domain experts with formal training in Bangla language and linguistics.

Disagreements were resolved through structured deliberation sessions in which annotators jointly examined the semantic, cultural, and contextual dimensions of the entry until full agreement was reached. An entry was finalized only when complete consensus was achieved.

Given this iterative consensus construction approach, formal agreement metrics such as Cohen's κ were not computed, as the final annotation reflects negotiated expert agreement rather than independently produced labels. This methodology prioritizes depth of linguistic interpretation and cultural nuance over statistical aggregation, which is particularly important for culturally grounded figurative expressions in low-resource settings.

3.4. Experimental Evaluation: Comprehensive LLM Benchmarking

To establish the dataset's utility as a benchmark resource, we conducted a comprehensive evaluation of 30 state-of-the-art LLMs selected through a strategic sampling methodology that ensures maximal representation across key dimensions of modern AI development.

Model Selection Strategy. Our model portfolio was constructed to capture the complete ecosystem

of contemporary language models through deliberate stratification across four critical axes:

- **Architectural Paradigms:** We included transformer variants (Llama, GPT), mixture-of-experts (Mixtral-8x22B, Mistral-Saba), and specialized architectures (Tongyi-DeepResearch, Claude-Haiku) to assess how structural differences affect cross-linguistic reasoning.

- **Scale Spectrum:** The selection spans five orders of magnitude, from compact 4.5B parameter models (Arcee-AI's AFM-4.5B) for edge deployment scenarios to massive 405B parameter models (Llama-3.1-405B) representing current scaling limits, enabling analysis of scale-performance relationships.

- **Geopolitical Distribution:** Our cohort includes models from the US (OpenAI, Anthropic, Meta), Chinese developers (Qwen, Baidu, Alibaba), European developers (Mistral), and others, capturing diverse training data compositions and cultural biases that may impact Bangla idiom comprehension.

- **Economic and Access Models:** We balanced proprietary production systems (GPT-4.1, Gemini-2.5) with open-source alternatives (Llama series, Gemma) and research-focused models (Skyfall-36B) to evaluate how development incentives and transparency affect performance.

This strategic sampling ensures our benchmark assesses not merely individual model capabilities but reveals broader patterns about how architectural choices, scaling laws, training data diversity, and development ecosystems influence cross-cultural linguistic understanding.

Experimental Design. We employed a zero-shot evaluation protocol on a subset of 100 culturally significant idioms. This subset was meticulously curated through iterative expert discussions by linguistics specialists who reviewed the entire *Bangla Bagdhara* dataset, explicitly avoiding any computational or random selection. Priority was given to idioms that are not only commonly used in natural language and daily conversation but also possess subtle or complex figurative contexts, ensuring the test set represents a meaningful spectrum of linguistic difficulty. The model cohort, intentionally including specialized variants (Tongyi-DeepResearch for academic contexts, MiniMax-M1 for multimodal grounding) alongside general-purpose models, was identically prompted to explain each idiom's meaning and cultural connotations.

Prompt Design. All models received identical prompts in Bangla, designed to elicit natural-language understanding from a native speaker's perspective. An example of the prompt is shown in Figure 3. As required for non-English content, the English translation of the core prompt given to the models is provided below:

> "Suppose you are an ordinary Bengali-speaking person from Bangladesh. You will be given an idiom/phrase below. You have to tell its meaning. If you see it in a sentence or writing, what meaning will you think of? In what sense is this word/phrase used in the Bangla language? [...] The most important thing is what you understand from the word/phrase, that's what we want to hear."

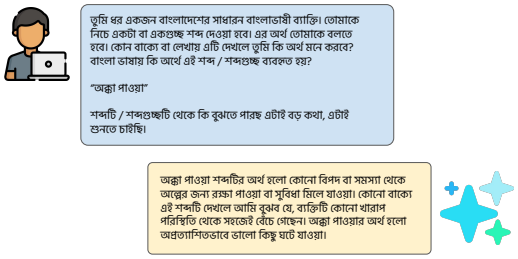

Figure 3: Example interaction with an LLM

Score	Description of Comprehension
5	Precise and complete figurative meaning
4	Correct figurative meaning
3	Correct meaning with minor inaccuracies
2	Vague/partial figurative grasp
1	Literal meaning only
0	Complete misunderstanding

Table 3: Evaluation rubric for semantic comprehension of figurative meaning.

Evaluation Protocol. An expert panel of five Bangla linguistics specialists conducted a double-blind review of all 3,000 responses. Each response was scored on a validated 6-point rubric assessing the semantic comprehension of the figurative meaning, detailed in Table 3. This rigorous protocol enables comparative analysis of model capabilities against human performance.

3.5. Human Baseline

To contextualize the LLM performance and establish a native-speaker benchmark, we administered the same 100-idiom question set to a cohort of 15

Bangladeshi students (6 high school and 9 undergraduate). The evaluation was conducted under controlled conditions, providing flexible time while prohibiting access to the internet or any additional textual resources.

This human cohort achieved a mean score of 417 out of 500 (83.4%). This translates to an average score of 4.17 per question on the 6-point evaluation rubric (see Table 3), with a standard deviation of 0.3936. This result serves as the human baseline for figurative and cultural comprehension against which the model performances are measured.

4. Dataset Description

The *Bangla Bagdhara* dataset comprises **10,361 annotated Bangla idioms**, making it the largest idiomatic dataset for Bangla. This section provides key statistics and characteristics.

4.1. Corpus Statistics and Semantic Coverage

The dataset contains **37,129 tag assignments** with an average of **3.58 semantic tags per idiom**. Importantly, semantic tags are not drawn from a predefined classification set, but are flexibly assigned according to the nuanced semantics and contextual meanings expressed in each idiom. This approach yielded **9,919 unique semantic tags**, demonstrating exceptional semantic diversity and capturing the full spectrum of idiomatic meanings in Bangla.

4.2. Sentiment and Domain Distribution

Sentiment analysis reveals a **negative bias** (47.1%, 4,882 idioms), with neutral at 37.4% (3,877) and positive at 15.5% (1,601), reflecting Bangla idioms' cautionary nature.

Domain coverage shows **conversation-related domains dominate** (53.1%, 5,498 idioms), with literature at 13.6% (1,409 idioms) and specialized domains (education, media) ensuring broad NLP applicability (Table 4).

Usage Domain	Percentage
Everyday Conversation	30.5%
Literature	13.6%
Daily Informal Talks	12.4%
Media	8.4%
Education	7.2%

Table 4: Primary Usage Domain Distribution

4.3. Geographical and Cultural Coverage

The corpus shows extensive geographical representation: **85.5% country-wide** (8,854 idioms), 11.2% rural (1,161), and 3.3% urban (346) (Figure 4). Cultural annotations reveal 99.7% have cultural significance, 13.5% carry religious significance, and 6.3% bear historical significance.

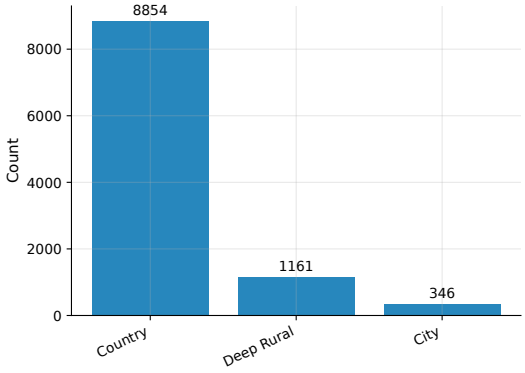

Figure 4: Dataset Geographical and Cultural Coverage

4.4. Frequency and Thematic Analysis

Frequency distribution balances accessibility and completeness: **75.1% common idioms** (7,777) and 17.2% rare (1,781), ensuring practical NLP applications while preserving linguistic diversity.

Thematic analysis reveals rich semantic patterns, with tags organically emerging from idiom meanings rather than being constrained by predefined categories. This flexible tagging approach captured four primary clusters: Emotional/Affective expressions, Conceptual/Literary devices, Social/Communicative functions, and Behavioral/Moral categories.

4.5. Register and Annotation Quality

Register distribution shows a **literary-to-conversational ratio of 1:4.6**. The dataset achieves **100% annotation completeness** for primary schema fields and 99.7% for cultural significance, ensuring reliable model training.

The comprehensive annotation framework, particularly the semantic tagging system that adapts to each idiom's unique meaning rather than imposing fixed categories, positions the *Bangla Bagdhara* dataset as a foundational resource for Bangla NLP in figurative language understanding and cultural computing.

5. Model Performance Analysis

The comprehensive evaluation of 30 state-of-the-art language models on the selected subset of the *Bangla Bagdhara* dataset benchmark reveals a striking and systemic failure in Bengali figurative language comprehension. A top-level overview of the

results, detailed in Table 5, shows that no model surpassed the 50% performance threshold. The leading model, `google/gemini-2.5-flash`, scored only 238 out of 500 (47.6%), while the majority of models exhibited catastrophic failure rates, often exceeding 80-90%. This poor performance is consistent across all model sizes, architectures, and geographic origins, pointing to a fundamental gap in current multilingual training paradigms rather than isolated model-specific weaknesses.

5.1. Performance Stratification and Variance

The results show a distinct stratification of model capabilities. A top tier emerges, led by `google/gemini-2.5-flash` (Mean Score: 2.38), followed by `anthropic/claude-3.5-sonnet` (1.93) and `deepseek/deepseek-chat-v3.1` (1.73). These models demonstrate a partial capacity for the task, achieving perfect scores (5/5) on 38, 30, and 23 idioms, respectively.

However, this potential is undermined by extreme performance inconsistency. All top-tier models exhibit high standard deviations (SD > 2.15) and Coefficients of Variation (CoV) ranging from 0.99 to 1.25. Such high variance suggests that their success on certain idioms is unpredictable and not indicative of a generalizable understanding of Bengali figurative language.

5.2. Prevalence of Task Failure

A critical finding is the pervasive rate of task failure, defined as a score of 0, across all models. Even the leading model, `gemini-2.5-flash`, failed completely on 46 of the 100 idioms. The failure rate escalates dramatically for lower-ranked models, with over half of the evaluated models failing on more than 80% of the test cases.

5.3. Analysis of Architectural and Training Patterns

A deeper analysis of the results reveals no correlation between performance and the models' underlying characteristics, such as size, architecture, or regional origin. This suggests the problem is universal to the current generation of LLMs.

Insignificance of Geographic Origin: High- and low-performing models originate from developers across the US, China, and Europe. For instance, top performers include models from Google (US) and DeepSeek (China), while poor performers also come from diverse geographic backgrounds. This lack of a regional pattern indicates that current multilingual pre-training strategies, regardless of their origin, are uniformly deficient in capturing the required cultural context for Bengali.

Impact of Model Size: The benchmark demonstrates that larger model size does not confer a significant advantage. For example, `meta-llama/llama-3.1-405b-instruct`, one of the largest models evaluated, ranked 16th with a mean score of only 0.68. In contrast, the much smaller and more efficient `google/gemini-2.5-flash` leads the benchmark. This disparity strongly suggests that simply scaling up parameters is an insufficient strategy for improving performance on culturally-specific, low-resource language tasks.

Architectural Paradigm: The rankings are populated by a mix of architectural designs, including dense transformer models and Mixture-of-Experts (MoE) architectures like `mistralai/mixtral-8x22b-instruct`. No single paradigm demonstrated a clear advantage; in fact, the aforementioned MoE model was one of the worst performers, scoring only 0.02. This implies that the performance deficit is not an architectural flaw but a data and training issue. Furthermore, specialized models like the research-oriented `alibaba/tongyi-deepresearch-30b` (rank 19) and the multimodal `qwen/qwen3-vl-8b-instruct` (rank 29) also failed to perform, confirming that neither academic specialization nor visual grounding currently compensates for this core linguistic gap.

6. Limitations and Future Directions

The development of the *Bangla Bagdhara* dataset provides a robust foundation for Bangla computational linguistics and illuminates several key areas for future enhancement. The current resource's limitations suggest a clear roadmap for advancing the field, where further work can be done in the following directions:

- **Temporal and Generational Expansion:** Future work can capture contemporary digital neologisms for diachronic study and develop generation-aware frameworks to analyze how idiom interpretation varies across demographics.
- **Geographical and Dialectical Representation:** Coverage can be deepened to include underrepresented dialects (e.g., Sylheti, Chittagonian) and urban linguistic innovations, facilitating robust dialect-aware systems.
- **Annotation Frameworks and Scalability:** The 19-field schema can be expanded using semi-supervised or community-sourced methods for efficient scaling, alongside developing more nuanced metrics for cultural significance.
- **Sentiment Analysis and Domain Adaptation:** The observed sentiment distribution

Rank	Model	Total	Mean	Std	Min	Max	IQR	CoV	Perfect	Fail
1	gemini-2.5-flash	238	2.38	2.36	0	5	5.0	0.99	38	46
2	claude-3.5-sonnet	193	1.93	2.25	0	5	5.0	1.17	30	53
3	deepseek-chat-v3.1	173	1.73	2.16	0	5	4.0	1.25	23	58
4	llama-4-maverick	161	1.61	2.16	0	5	4.0	1.34	22	61
5	gemini-2.5-flash-lite	145	1.45	2.18	0	5	4.0	1.50	24	67
6	chatgpt-4o-latest	141	1.41	2.11	0	5	4.0	1.50	19	66
7	gpt-4.1	136	1.36	2.02	0	5	3.0	1.49	18	64
8	qwen3-max	128	1.28	2.00	0	5	3.0	1.56	15	69
9	claude-haiku-4.5	109	1.09	1.85	0	5	2.0	1.70	10	72
10	llama-4-scout	108	1.08	1.85	0	5	2.0	1.71	12	71
11	mistral-medium-3.1	86	0.86	1.72	0	5	0.0	2.00	11	76
12	grok-3-mini	84	0.84	1.70	0	5	0.0	2.02	10	77
13	gemma-3n-e4b-it	83	0.83	1.60	0	5	1.0	1.92	7	74
14	gemma-3-12b-it	81	0.81	1.59	0	5	0.0	1.96	6	76
15	kimi-k2-0905	78	0.78	1.69	0	5	0.0	2.16	10	80
16	llama-3.1-405b	68	0.68	1.58	0	5	0.0	2.33	8	83
17	grok-4-fast	66	0.66	1.54	0	5	0.0	2.33	8	82
18	nova-pro-v1	49	0.49	1.34	0	5	0.0	2.74	4	87
19	tongyi-deepresearch	48	0.48	1.28	0	5	0.0	2.67	5	84
20	minimax-m1	46	0.46	1.36	0	5	0.0	2.95	6	88
21	llama-3.3-70b	36	0.36	1.02	0	5	0.0	2.83	3	85
22	glm-4.5	34	0.34	1.14	0	5	0.0	3.35	3	91
23	ernie-4.5-21b	33	0.33	0.99	0	5	0.0	2.99	1	88
24	longcat-flash-chat	30	0.30	0.90	0	4	0.0	3.02	0	88
25	mistral-saba	22	0.22	0.81	0	5	0.0	3.69	1	92
26	cydonia-24b-v4.1	19	0.19	0.75	0	4	0.0	3.94	0	93
27	skyfall-36b-v2	4	0.04	0.24	0	2	0.0	6.07	0	97
28	mixtral-8x22b	2	0.02	0.20	0	2	0.0	10.00	0	99
29	afm-4.5b	0	0.00	0.00	0	0	0.0	0.00	0	100
29	qwen3-vl-8b	0	0.00	0.00	0	0	0.0	0.00	0	100

Table 5: Model performance comparison across various metrics.

(47.1% negative) offers a distinct opportunity to develop robust domain adaptation techniques for sentiment analysis models.

- **Benchmark Development:** A significant opportunity exists to establish standardized benchmarks and evaluation tasks (detection, interpretation, generation) to systematize research in Bangla idiom processing.
- **Benchmark Scope and Evaluation Robustness:** The 100-idiom test set can be significantly expanded to ensure robustness, while objective metrics can be investigated to complement subjective, consensus-based expert evaluations.
- **Broadening Model Diversity in Benchmarking:** The benchmark can be continually updated with a wider array of models, new architectures, and training paradigms beyond the current 30 to better track state-of-the-art capabilities.
- **Computational Efficiency and Accessibility:** Lightweight access methods, specialized data subsets, and APIs can be developed to support researchers in resource-constrained environments.
- **Annotation Rigor:** Future annotation efforts provide an opportunity to compute formal agreement metrics (e.g., Cohen's κ) through parallel annotation, complementing the current qualitative consensus method.

In summary, each identified limitation of the *Bangla Bagdhara* dataset presents a clear opportunity not only to enhance Bangla figurative language resources but also to advance NLP methodologies for low-resource languages globally.

7. Ethics Statement

This work involves the construction and annotation of a culturally grounded idiom dataset in Bangla and the evaluation of language models on a figurative language understanding task. The dataset consists exclusively of publicly documented idiomatic

expressions collected from authoritative literary and archival sources. No personally identifiable information (PII) or private data were collected or included.

Human evaluation was conducted with 15 volunteer participants (high school and undergraduate students) who are native Bangla speakers from Bangladesh. Participation was voluntary, no sensitive personal data were recorded, and no compensation-based coercion was involved. The evaluation task consisted solely of interpreting idiomatic expressions and did not expose participants to harmful or sensitive content.

The annotation process was carried out by domain experts with formal training in Bangla language and linguistics. Care was taken to preserve cultural authenticity while avoiding offensive or discriminatory reinterpretations of idioms. As idiomatic expressions may reflect historical or socio-cultural biases, the dataset is intended strictly for research and educational use in computational linguistics.

The evaluation of large language models was conducted through a double-blind protocol to ensure impartial scoring. The authors acknowledge that culturally grounded datasets may encode societal biases present in historical language usage, and future work will continue to examine responsible and context-aware deployment.

8. Conclusion

As the first large-scale corpus of its kind, the *Bangla Bagdhara* dataset contains 10,361 idioms annotated with a detailed 19-field schema designed to capture their rich semantic, pragmatic, and cultural nuances. This resource addresses a long-standing gap in figurative language analysis for low-resource languages.

Our benchmark evaluations reveal a significant performance gap in state-of-the-art large language models, none of which surpassed 50% accuracy on Bangla idiom comprehension. This finding highlights the urgent need for language models that are more deeply attuned to cultural and figurative contexts.

The development of the *Bangla Bagdhara* dataset is an ongoing initiative. We are committed to continuously expanding and refining the corpus to enhance its utility for diverse NLP applications, such as machine translation, cultural analytics, and educational technology. By making this resource publicly available, we aim to stimulate innovation in Bangla NLP and encourage the creation of similar culturally grounded resources for other underrepresented languages, contributing to a more inclusive and linguistically aware AI ecosystem.

9. Bibliographical References

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Sharma. 2018. No more beating about the bush : A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Paul Baker, Andrew Hardie, Tony McEnery, and B. D. Jayaram. 2003. Constructing corpora of south asian languages.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tanmoy Chakraborty, Dipankar Das, and Sivaji Bandyopadhyay. 2014. Identifying bengali multiword expressions using semantic clustering.

Ashwini Deo and Devyani Sharma. 2006. Typological variation in the ergative morphology of Indo-Aryan languages. Final draft, to appear in *Linguistic Typology*.

Michael Flor, Xinyi Liu, and Anna Feldman. 2025. A survey of idiom datasets for psycholinguistic and computational research. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Long and Short Papers*, pages 90–100, Hannover, Germany. HsH Applied Academics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Ayesha Khatun, Md Gulzar Hussain, Md Islam, Sumaiya Kabir, and Md Mahin. 2020. An emper-

ical framework of idioms translator from bengali to english: Rule based approach.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2025. Comparative study of multilingual idioms and similes in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8680–8698, Abu Dhabi, UAE. Association for Computational Linguistics.

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. 2018. LIdioms: A multilingual linked idioms data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Geetanjali Rakshit and Jeffrey Flanigan. 2025. Alankaar: A dataset for figurativeness understanding in bangla. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 998–1002, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Naziya Mahamdul Shaikh, Jyoti D. Pawar, and Mubarak Banu Sayed. 2024. Konidioms corpus: A dataset of idioms in Konkani language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9932–9940, Torino, Italia. ELRA and ICCL.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Laura Street, Nathan Michalov, Rachel Silverstein, Michael Reynolds, Lurdes Ruela, Felicia Flowers, Angela Talucci, Priscilla Pereira, Gabriella Morgon, Samantha Siegel, Marci Barousse, Antequa Anderson, Tashom Carroll, and Anna Feldman. 2010. Like finding a needle in a haystack: Annotating the American national corpus for idiomatic expressions. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).