International Journal of Informatics and Communication Technology (IJ-ICT)
Vol. 15, No. 2, June 2026, pp. 823~838
ISSN: 2252-8776, DOI: 10.11591/ijict.v15i2.pp823-838

Can machines imagine? Critical thinking and cultural reasoning in multimodal-multilingual AI

Mohammad Awad AlAfnan[1], Siti Fatimah MohdZuki[2], Shefa Mohammad AlAfnan[3]
[1]Department of Liberal Arts, American University of the Middle East, Egaila, Kuwait
[2]University Technology MARA, Shah Alam, Malaysia
[3]International Islamic University Malaysia, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Jul 23, 2025
Revised Mar 30, 2026
Accepted Apr 15, 2026

Keywords:

Cross-cultural communication
Ethical AI
Low-resource languages
Multilingual AI
Multimodal language models

ABSTRACT

Effective communication across languages and cultures is essential in today's interconnected world. Multimodal-multilingual language models (MMMLMs) aim to advance this goal by integrating text, speech, and visual understanding across diverse linguistic contexts. This study evaluates four leading MMMLMs-GIT, mPLUG, CLIP, and Whisper + GPT-4V-on cross-lingual and cross-modal tasks, including image captioning, visual question answering, speech-to-image generation, and idiomatic translation. Performance was assessed in high-resource (English, Arabic), medium-resource (Malay), and low-resource (Macedonian) settings. Results show strong performance in structured tasks but notable limitations in cultural reasoning, figurative language interpretation, and semantic grounding in low-resource environments. GIT delivered the most consistent multilingual results, while Whisper + GPT-4V excelled in fluency yet lacked cultural sensitivity. To address these gaps, the study proposes culturally informed evaluation protocols that integrate quantitative metrics such as BLEU, CIDEr, and F1 with qualitative, community-centered approaches. These include cross-cultural annotation panels, inter-rater reliability validation using Cohen's kappa, and a novel "cultural fidelity" metric to measure alignment with culturally specific norms. The findings emphasize the need for inclusive datasets, ethical development, and interdisciplinary collaboration to ensure MMMLMs support equitable and culturally aware global communication.

Corresponding Author:

Mohammad Awad AlAfnan
Department of Liberal Arts, American University of the Middle East
Block 6, Building 1, Egaila, Kuwait
Email: Mohammad.al-afnan@aum.edu.kw

1. INTRODUCTION

In an era of globalization and digital transformation, communication across linguistic and cultural boundaries has become essential to international collaboration, economic development, and social cohesion [1]. However, despite significant technological advances, linguistic diversity and the proliferation of communication modalities continue to present substantial challenges to mutual understanding. Traditional language processing systems, while effective within specific linguistic or modal confines, often fall short in addressing the complex demands of real-world multilingual and multimodal communication [2]. This gap underscores the pressing need for integrated approaches that transcend monolingual and monomodal paradigms. One of the most promising developments in this regard is the emergence of multimodal-multilingual language models (MMMLMs), AI systems designed to interpret and generate content across multiple human languages and communicative modalities [3].

MMMLMs represent a novel convergence of two previously distinct areas of artificial intelligence: multilingualism, which enables language models to process and produce multiple languages, and multimodality, which facilitates the understanding and generation of content across various data types, including text, speech, images, and video [4]. While significant progress has been made in each area independently, their integration has only recently become feasible due to breakthroughs in model architecture, pretraining strategies, and the availability of large, diverse datasets. The resulting models offer unprecedented potential for cross-cultural communication, enabling text-based translation and holistic, context-sensitive interpretation of meaning in real-world, multimedia environments [5].

The importance of such capabilities is increasingly evident across multiple sectors. In international diplomacy, for example, MMMLMs could facilitate more accurate and culturally nuanced translations during negotiations, reducing the risk of misinterpretation [6]. In global health, these models can support disseminating equitable information through localized visual and audio content, particularly critical in crisis contexts, such as pandemics or natural disasters [7]. Similarly, in education, MMMLMs can promote inclusive and accessible learning by translating multimodal instructional materials, such as infographics, video lectures, and interactive simulations, into multiple languages and formats that cater to diverse cognitive and linguistic needs [8]. MMMLMs can enhance user experience and business and customer service engagement by localizing multimodal interfaces and communications in real time [9].

Technologically, MMMLMs build on recent advances in large-scale transformer architectures, such as GPT, BERT, and vision transformers (ViTs), which enable models to process high-dimensional inputs from diverse sources [10]. Training such models requires vast data and sophisticated strategies for cross-modal alignment and multilingual tokenization. For instance, aligning an image with captions in different languages or synchronizing spoken dialogue with corresponding visual scenes entails the creation of joint embedding spaces that preserve semantic equivalence across languages and modalities. Moreover, model architectures must be able to disambiguate contextual meaning, especially in settings where a single gesture, word, or visual cue may carry culture-specific significance [11].

While MMMLMs offer promising solutions to longstanding communication barriers, their development is not without significant challenges. First, there is the issue of data imbalance. High-resource languages, such as English, Mandarin, and Spanish, dominate available training corpora, while low-resource and endangered languages are often underrepresented or excluded entirely. This not only perpetuates existing inequities but also limits the applicability of these models in diverse global contexts. Efforts to address this issue have included transfer learning, synthetic data generation, and community-based data collection, yet these approaches remain resource-intensive and require careful ethical oversight.

Second, the computational and environmental costs of training MMMLMs are substantial [12]. Fusing multiple modalities and languages significantly increases the model's size and complexity, demanding immense processing power and energy consumption. This raises important questions regarding the sustainability and scalability of these systems, particularly when deployed in regions with limited infrastructure. Research into more efficient architectures, such as sparse transformers and model distillation techniques, is ongoing and represents a critical direction for future work.

Third, and perhaps most critically, ethical and sociocultural implications must be considered. MMMLMs, like all machine learning systems, reflect the biases and limitations of their training data [13]. When these biases pertain to language, culture, gender, or ethnicity, the consequences can be especially harmful. For example, a model trained on biased image-text pairs might perpetuate stereotypes in visual storytelling. At the same time, inaccuracies in the translation of health-related content could lead to real-world harm. Therefore, the design and deployment of MMMLMs must be guided by algorithmic fairness, linguistic equity, and cultural sensitivity principles. Interdisciplinary collaboration, with linguists, anthropologists, ethicists, and community stakeholders, is essential to ensure that these systems are both technologically robust and socially responsible.

MMMLMs also raise complex questions regarding the nature of meaning itself [14], [15]. While traditional translation focuses on linguistic equivalence, MMMLMs convey multimodal meaning across cultural and linguistic boundaries. This involves lexical translation and the accurate interpretation of non-verbal cues such as facial expressions, body language, tone of voice, and cultural symbolism. These elements are often deeply embedded in local contexts and may not have direct equivalents across languages or cultures. Therefore, MMMLMs must be equipped with pragmatic and cultural competence, which remains a formidable challenge in computational modeling.

Despite these limitations, early implementations of MMMLMs have demonstrated encouraging results [16]. Multilingual, multimodal, and multitask pretraining (M3P) and generative image-to-text transformer (GIT) have shown the feasibility of integrating multilingual and multimodal capabilities within a single framework. Commercial models like GPT-4V, Gemini, and contrastive language-image pretraining (CLIP) have begun to incorporate image, text, and speech understanding across multiple languages,

setting new benchmarks for performance. However, comprehensive evaluations, particularly in real-world, low-resource, or high-stakes settings, remain limited and represent a fertile area for further investigation.

The integration of multimodal and multilingual processing thus signals a transformative moment in the evolution of language technologies [17]. MMMLMs not only promise more fluid and inclusive communication across borders but also challenge existing paradigms in artificial intelligence, linguistics, and global development. As such, they invite a re-examination of what it means to "understand" and "translate" in a multilingual, multimedia world. Developing these models is not merely a technical challenge, but a cultural and ethical project that must center on human values, linguistic diversity, and equitable access at every stage [18].

In light of these considerations, this article aims to comprehensively explore MMMLMs as a foundational technology for global communication [19]. It begins by outlining the theoretical and architectural underpinnings of multimodal and multilingual systems, then surveys current models and datasets. Subsequently, it examines key challenges, ranging from computational scalability to cultural representation, and proposes a framework for evaluating the performance and social impact of MMMLMs. Ultimately, it provides recommendations for interdisciplinary research and policy development to ensure that these technologies serve the broadest possible range of users equitably and responsibly.

2. LITERATURE REVIEW

The natural language processing (NLP) field has undergone a significant transformation over the past decade, primarily driven by the development of large-scale, pre-trained models [20], [21]. These models, such as BERT [22], GPT-2 [23], and GPT-3 [24], have demonstrated remarkable performance across a wide range of language tasks. Initially designed for monolingual tasks in English, they have since evolved into multilingual models, capable of processing and generating text in over 100 languages. The development of multilingual models, such as mBERT (multilingual BERT) and XLM-R [25], has been pivotal in expanding the inclusivity of language technologies. However, these models primarily focus on text-based input and output, and do not natively handle other modes of communication such as images, speech, or video.

Parallel to advancements in multilingual modeling, multimodal learning has gained significant momentum. Models such as CLIP [26] and DALL·E [27] have demonstrated the power of combining visual and textual representations in a unified embedding space. These models leverage contrastive learning or autoregressive generation to associate image features with natural language descriptions, enabling applications in image captioning, visual question answering (VQA), and zero-shot classification. Similarly, speech-text integration has advanced through models like Whisper [28], which provide high-quality transcription and translation capabilities. Despite these successes, early multimodal models often operated in monolingual environments and could not generalize across languages.

The convergence of these two strands, multimodal and multilingual modeling, has only recently begun to materialize in MMMLMs. One of the earliest attempts to explore this intersection was the Unicoder-VL [29] and M3P [30], which combined multilingual text representations with visual understanding. These models were pre-trained on image-text pairs from multiple languages, enabling cross-lingual image retrieval and captioning across various languages. M3P in particular demonstrated that pretraining across both modalities and languages leads to performance gains in multilingual vision-language tasks. However, such models were limited by the availability of parallel multimodal data, particularly for low-resource languages.

Further advancements were marked by developing models such as mPLUG [31] and GIT [32], which integrated image and text modalities, focusing on multilingual datasets. These models introduced transformer-based architectures capable of simultaneously handling complex interactions between modalities and languages [33]. GIT, for example, leveraged a generative framework that enabled tasks such as multilingual image captioning and cross-lingual visual storytelling. Using shared token spaces and language-agnostic representations improved model generalization across culturally diverse datasets. Despite this, the performance of MMMLMs remains skewed towards high-resource languages and widely available visual content, pointing to ongoing issues of representational inequality [34].

Recent contributions have also emphasized incorporating speech and video into the multimodal-multilingual pipeline. The work of Chen *et al.* [35], which utilizes VideoLLM and integrates Whisper for multilingual automatic speech recognition, represents a significant step toward inclusive multimedia understanding. These systems can process spoken input in multiple languages, align it with video frames, and generate coherent summaries or responses in the target language. Such capabilities are vital for applications in education, accessibility, and media localization. However, the training of such models is constrained by the scarcity of aligned multilingual audiovisual datasets and the complexity of integrating temporally dependent data, such as speech and video, with static modalities like text or images.

A vital aspect explored in the literature is the nature of cross-modal and cross-lingual alignment. Studies such as those by AlAfnan [36] and Yang *et al.* [37] have highlighted the difficulty of creating both language-agnostic and modality-agnostic embeddings. These embeddings must preserve semantic relationships across input types, such as visual cues and spoken words, while remaining coherent across languages with diverse grammatical and syntactic structures. Contrastive learning, self-supervised pretraining, and masked modeling techniques have been widely employed to address these challenges [38]. Nonetheless, aligning semantics across modalities and languages in a culturally sensitive manner remains a core research bottleneck.

Beyond technical considerations, a growing body of scholarship has examined the social, ethical, and cultural implications of MMMLMs. Bender *et al.* [39] have argued for greater attention to the socio-linguistic context of model training and deployment, cautioning against language technologies' "colonial" tendencies that prioritize dominant languages and Western cultural frameworks. Research by Joshi *et al.* [40] on low-resource languages reinforces this concern, pointing out that the vast majority of existing NLP systems are optimized for only a handful of major languages, without intentional efforts to include marginalized languages and communication styles, such as sign languages or indigenous oral traditions. MMMLMs risk exacerbating global digital divides.

The literature raises essential questions about cultural representation and the translation of meaning. Communicative elements such as idioms, gestures, visual symbolism, and emotional tone are often deeply embedded in artistic contexts. Pavlick [41], even state-of-the-art models struggle to translate or generate content requiring cultural inference accurately. This is particularly problematic in multimodal contexts, where a single emoji, gesture, or visual reference may carry vastly different meanings depending on the audience. MMMLMs must go beyond literal translation and develop pragmatic competence, a challenge that demands interdisciplinary collaboration with fields such as anthropology, semiotics, and cognitive science.

Evaluation metrics for MMMLMs are another area receiving increased scholarly attention. Traditional NLP benchmarks such as BLEU or ROUGE scores are insufficient for capturing performance across multiple modalities and languages. New multimodal evaluation frameworks, such as VQA-X [42] and Multi30K [43], have been proposed to assess cross-lingual and visual understanding. However, these datasets often have limited linguistic diversity and are heavily skewed toward Western cultural references. Calls for more inclusive and culturally diverse benchmarks are growing, with researchers emphasizing the importance of user-centered evaluations that reflect real-world use cases and multilingual populations [44].

The role of community-driven data collection and annotation is emerging as a best practice in this context. Initiatives such as masakhane, common voice, and data for indigenous languages have demonstrated that involving local speakers and community experts in developing multilingual datasets yields more accurate and ethically sound outcomes. Such approaches are especially critical in developing MMMLMs, where the semantic alignment of modalities often depends on culturally contextualized annotation.

Ultimately, research on the scalability and deployment of MMMLMs in low-resource environments remains in its infancy [45], [46]. While most models are trained and tested in high-performance computing environments, their application in real-world, bandwidth-constrained, and low-literacy contexts remains largely unexplored. Lightweight models, edge computing, and language-specific fine-tuning are being investigated as potential solutions [47]; however, significant work remains to ensure that MMMLMs can operate reliably and equitably outside laboratory conditions.

As such, the existing literature provides a robust foundation for developing MMMLMs, with significant strides made in architecture, training strategies, and initial applications. Nonetheless, persistent challenges relate to data diversity, cultural representation, ethical deployment, and evaluation. Bridging these gaps requires sustained interdisciplinary research, inclusive data practices, and policy frameworks prioritizing linguistic justice and digital equity. As the field progresses, MMMLMs must evolve not only in response to technological innovations but also as sociotechnical systems grounded in the plural realities of global communication.

3. METHOD

This study adopts a mixed-methods research design to examine the capabilities, limitations, and potential of MMMLMs in facilitating global communication. The methodology comprises three primary components: model selection and benchmarking, dataset preparation and evaluation protocols, qualitative content analysis, and cultural contextualization. This comprehensive approach enables technical assessment and sociolinguistic critique, ensuring that the study captures performance metrics and interpretive depth across diverse linguistic and cultural contexts.

3.1. Model selection and benchmarking

The first phase of this study involved selecting representative MMMLMs that exemplify the current state of the art in multimodal and multilingual capabilities. The models selected for evaluation include:

- CLIP: primarily for vision-language alignment.
- GIT: a generative multimodal model with multilingual capabilities.
- mPLUG: a multilingual, multimodal transformer model that integrates image, text, and limited speech.
- Whisper + GPT-4V pipeline: a combined speech-to-text and visual-language model chain enabling end-to-end multimodal-multilingual processing.

These models were selected based on criteria such as architectural diversity, multilingual capabilities**,** support for multiple modalities (text, image, and/or speech), and accessibility for evaluation (i.e., open-source or research-use availability). CLIP and Whisper + GPT-4V were included in the study despite lacking full multimodal-multilingual (MMMLM) integration because they represent crucial partial architectures that advance two complementary axes of multimodal AI-vision-language and speech-language alignment. CLIP demonstrates exceptional image-text contrastive learning, establishing a shared embedding space where visual and linguistic semantics correlate, making it a foundational model for visual grounding. Whisper + GPT-4V, in contrast, forms a speech-text-vision pipeline through modular coupling: Whisper converts multilingual speech into text, while GPT-4V interprets that text alongside visual input, yielding near-end-to-end multimodal reasoning. Although neither model natively performs joint multilingual and multimodal training (as GIT or mPLUG do), their inclusion highlights the field's transitional phase from unimodal specialization to integrated MMMLMs. Architecturally, GIT, and mPLUG employ unified transformer backbones trained jointly on cross-lingual image-text pairs, enabling shared token spaces and cross-modal attention; CLIP and Whisper + GPT-4V, by contrast, rely on modular alignment-separate encoders bridged through embeddings or chained inference. Thus, their selection underscores the methodological diversity in the evolution toward complete MMMLM systems and illustrates how modular models contribute essential building blocks for future integrated, culturally aware architectures.

Evaluating proprietary models such as GPT-4V through API access presents significant methodological and ethical limitations. First, API-based evaluation limits transparency, as researchers cannot inspect the model architecture, the composition of the training data, or the fine-tuning strategies, making it challenging to attribute performance outcomes to specific design factors. This opacity limits reproducibility and undermines scientific rigor, as results cannot be independently verified or replicated under identical conditions. Second, API interfaces often impose rate limits, content filters, and usage constraints that limit experimental control, particularly in multilingual or culturally sensitive tasks that require iterative testing. Third, the black-box nature of these models prevents fine-grained error analysis-researchers can observe outputs but not the internal decision pathways or latent embeddings that produced them, complicating the study of bias, cultural misrepresentation, or hallucination. Furthermore, API access typically depends on corporate policies and subscription tiers, raising questions of accessibility, equity, and long-term research continuity. Consequently, while APIs enable convenient benchmarking of cutting-edge models, they simultaneously obscure critical technical and ethical dimensions necessary for transparent and accountable evaluation in multimodal-multilingual AI research.

Each model was deployed within a controlled experimental environment using NVIDIA A100 GPUs, with code run through Python (PyTorch) frameworks. Pretrained weights were retained to simulate real-world usage rather than task-specific fine-tuning, which may not be feasible in low-resource settings. This setup allowed for comparing each model's performance across cross-lingual image captioning, multilingual VQA, speech-to-image narrative generation**,** and translation of multimodal context tasks.

3.2. Dataset preparation and evaluation protocols

To evaluate the effectiveness of these MMMLMs, we assembled a collection of multilingual and multimodal datasets encompassing a wide range of languages, domains, and cultural backgrounds. These datasets were selected to represent both high-resource and low-resource language contexts, considering factors such as linguistic typology, geographical distribution, and multimodal diversity. The primary datasets consist of:

- Multi30K: multilingual image captions.
- Multicultural reasoning over vision and language (MaRVL): culturally grounded VQA dataset.
- Common voice + Flickr30k hybrid: a synthetic dataset combining Mozilla's multilingual speech corpus with image annotations.
- Wikipedia image-text (WIT): cross-lingual image-text pairs extracted from Wikipedia, available in over 100 languages.

Bilingual annotators verified translation quality as needed to maintain semantic accuracy throughout the samples. Moreover, an additional dataset featuring culturally relevant images and idioms was assembled

to evaluate the models' interpretive skills in less-represented contexts. This dataset encompasses traditional artifacts, culturally significant gestures, and localized metaphors annotated in Arabic, Malay, and Macedonian. Although Macedonian was categorized as a low-resource language in this study due to its limited digital representation and the scarcity of annotated multimodal datasets, it is essential to acknowledge that, linguistically, it belongs to the Indo-European family, specifically the South Slavic branch. This classification means it shares structural and lexical affinities with higher-resource languages such as Bulgarian and Serbian, as well as broader typological connections to other Indo-European languages. However, despite these linguistic proximities, Macedonian remains underrepresented in large-scale AI corpora and multimodal datasets, resulting in poorer model performance and limited semantic alignment. This disparity underscores that low-resource status in AI research is determined less by linguistic lineage than by data availability, corpus size, and inclusion in global training pipelines—factors that continue to disadvantage languages like Macedonian despite their rich linguistic heritage and established grammatical frameworks. For each task, standard and adapted evaluation metrics were employed. These include:

- BLEU, METEOR, and CIDEr for text-based captioning tasks.
- Accuracy and F1-score for VQA and classification tasks.
- Manual scoring protocols for culturally contextual narrative interpretation and semantic coherence.

To complement these quantitative metrics, we conducted human evaluations with bilingual speakers and cultural informants, who scored outputs on a Likert scale (1-5) for cultural appropriateness, fluency, and pragmatic adequacy. To ensure scoring consistency, inter-rater reliability was calculated using Cohen's kappa. To mitigate bias in human annotations during qualitative scoring, the study implemented several methodological safeguards to ensure consistency, transparency, and cultural sensitivity. First, all annotators were bilingual or multilingual speakers with cultural familiarity relevant to the languages and contexts under evaluation. Prior to data scoring, they underwent calibration sessions using sample outputs to establish shared interpretive criteria for assessing fluency, semantic alignment, and cultural relevance. A detailed annotation guide was provided to minimize subjective variation and clarify evaluative dimensions such as pragmatic accuracy and idiomatic equivalence. To quantify reliability across raters, inter-rater agreement was calculated using Cohen's kappa, ensuring statistical validation of consistency in qualitative judgments. Discrepancies in scoring were resolved through consensus discussions moderated by a linguistic expert, preventing dominance by any single annotator's perspective. Additionally, evaluators were instructed to document reasoning for outlier scores to maintain auditability and reflexivity in their interpretive decisions. These combined measures enhanced the validity and fairness of human evaluations while reducing potential cultural or cognitive bias in assessing multimodal-multilingual model outputs.

3.3. Qualitative content analysis and cultural contextualization

Although quantitative benchmarking plays a crucial role, it fails to fully address the complexities of how meaning is constructed across different cultures and modalities. Hence, this study employed qualitative content analysis to investigate MMMLMs' results thoroughly. This phase centered on the:

- Misinterpretation patterns: instances where the model produced semantically plausible but culturally inaccurate responses.
- Bias reflection: outputs that exhibited gender, racial, or cultural stereotypes.
- Pragmatic ambiguity: cases where translated meanings or generated captions failed to align with the context due to modal or linguistic shifts.

Each output was analyzed using an inductive method rooted in critical discourse analysis (CDA) and semiotic examination. Themes emerged from consistent patterns of interpretive errors or significant cultural omissions. The aim was to identify systematic deficiencies in model training and reasoning, particularly in low-resource or marginalized contexts.

An advisory group comprising linguists, anthropologists, and AI ethicists provided ongoing feedback during this phase. Their insights were instrumental in refining the annotation schema and identifying instances of symbolic or inferential errors that traditional NLP metrics did not capture.

3.4. Ethical framework and model of accountability

The study employed an ethics-by-design framework, adhering to guidelines established by the Montreal AI Ethics Institute and UNESCO's recommendation on the ethics of artificial intelligence [48].

Specific measures included:

- Ensuring language diversity and data transparency.
- Avoid using sensitive or personally identifiable content in model prompts or outputs.
- Including human-centered assessment in all evaluation phases to better reflect real-world communicative needs.

Bias assessments were carried out using automated tools, such as fairness indicators, and manual reviews. Additionally, models were evaluated for hallucination occurrences, specifically instances where the produced content included incorrect or fabricated information, particularly in multimodal translation and summarization tasks.

3.5. Limitations and scope

While the methodology aims to provide a holistic evaluation of MMMLMs, it is constrained by several factors:

- Dataset coverage: despite efforts to include diverse languages, accurate global coverage remains limited by data availability and annotation resources.
- Model access: proprietary models such as Gemini or GPT-4V with full multimodal capabilities were evaluated only to the extent allowed by public APIs.
- Generalization limits: the study focuses on zero-shot or few-shot performance and does not explore extensive fine-tuning or domain adaptation.

Future research should extend this methodology to sign language modalities**,** cognitive accessibility, and emergency communication scenarios, where MMMLMs can offer unique value if adapted appropriately.

4. RESULTS

The MMMLMs were evaluated across four core tasks: image captioning, VQA, speech-to-image generation, and cultural idiom translation. Results were analyzed quantitatively (using standard metrics) and qualitatively (using human evaluations and error analyses).

Using the Multi30 K and WIT datasets, multilingual image captioning was evaluated across four languages: English, Arabic, Malay, and Macedonian. The models assessed included CLIP, mPLUG, and GIT. Examples 1, 2, 3, and 4 illustrate the models' performance in evaluating multilingual image captions.

Image: A woman in traditional attire serving food to children at a cultural event.

Language	CLIP	mPLUG	GIT
EX1: English	**Caption**: A woman gives food to kids. **Analysis**: Factually correct, but lacks cultural nuance or descriptive richness. Very literal.	**Caption**: A woman in traditional attire serves food to children at a cultural celebration. **Analysis**: Improved detail and cultural context; includes "traditional clothes" and "celebration".	**Caption**: A smiling woman dressed in traditional festival attire serves a group of children during a community cultural event. **Analysis**: Most detailed and contextually rich. Captures emotion, attire, and the nature of the event.
EX2: Arabic	**Caption**: امرأة تعطي طعامًا للأطفال. (A woman gives food to children.) **Analysis**: Basic and correct, but lacks context or cultural elements.	**Caption**: امرأة ترتدي ملابس تقليدية تقدم الطعام للأطفال في احتفال ثقافي. (A woman wearing traditional clothes serves food to children at a cultural celebration.) **Analysis**: Similar to English, it adds cultural details.	**Caption**: " امرأة تبتسم ترتدي زيًا تقليديًا وتقدم الطعام للأطفال خلال مهرجان مجتمعي. (A smiling woman wears traditional attire and serves food to children during a community festival.) **Analysis**: Rich, contextually appropriate, emotionally expressive.
EX3: Malay	***Caption:*** Seorang wanita memberi makanan kepada kanak-kanak." *(A woman gives food to children)* *Analysis:* Accurate but lacks cultural specificity.	**Caption**: Seorang wanita berbaju tradisional sedang menghidang makanan kepada kanak-kanak di majlis kebudayaan. (A woman in traditional clothes is serving food to children at a cultural event) **Analysis**: Adds event and attire details.	***Caption:*** Seorang wanita yang tersenyum mengenakan pakaian tradisional sedang menghidangkan makanan kepada sekumpulan kanak-kanak dalam majlis komuniti. *(A smiling woman in traditional attire serves food to a group of children during a community gathering.)* ***Analysis:*** Full context captured, nuanced, and natural in tone.
EX4: Macedonian	***Caption:*** Жена дава храна на деца. *(A woman gives food to children)* ***Analysis:*** Grammatically correct but lacks any cultural insight.	**Caption**: Жена во традиционална облека служи храна на деца на културна прослава. (A woman in traditional clothes serves food to children at a cultural celebration.) **Analysis**: Captures cultural context but remains formulaic.	**Caption**: Насмеана жена во традиционална носија им служи храна на деца за време на заедничка културна прослава. (A smiling woman in traditional costume serves food to children during a community cultural celebration.) **Analysis**: Most fluent and culturally appropriate, reflects event atmosphere.

As Table 1 shows, GIT consistently outperformed the other models across all four languages. Macedonian scored lowest overall scores, highlighting continued performance gaps for low-resource languages.

Table 1. Captioning performance across languages (CIDEr score)

Model	English	Arabic	Malay	Macedonian
CLIP	1.11	0.88	0.85	0.73
mPLUG	1.22	1.01	1.00	0.91
GIT	1.26	1.10	1.03	0.95

VQA tasks represent a sophisticated intersection of computer vision and natural language processing. They utilize culturally contextual images derived from the MaRVL dataset. This dataset is specifically curated to include various images that reflect various cultural elements, making it a vital resource for understanding visual content through different cultural lenses. The results for VQA tasks are provided in Table 2.

Table 2. VQA task - accuracy (%)

Language	CLIP (%)	mPLUG (%)	GIT (%)
English	77.2	81.4	85.6
Arabic	71.3	76.0	80.1
Malay	68.9	74.5	79.0
Macedonian	63.2	70.2	75.6

Completing VQA tasks requires the model to interpret and analyze the images deeply, enabling it to answer questions inherently tied to the visual content presented. The scope of the questions ranges widely, covering various aspects of the images, including identifying objects, understanding actions, and grasping the cultural significance behind specific visual elements.

These tasks aim to enhance the model's ability to comprehend and appreciate the nuances of diverse cultural backgrounds reflected in the images. By doing so, the model is better equipped to provide accurate and contextually aware answers, considering the specific cultural contexts associated with each image. This becomes particularly significant when the goal is to answer questions in five target languages, as it requires linguistic proficiency and an awareness of cultural subtleties and variations across different contexts. Thus, VQA tasks are critical to developing linguistically adept and culturally intelligent models, ultimately promoting a more inclusive approach to visual understanding.

All models exhibited a notable decline in performance when applied to languages with lower resources (see Table 2). Despite the generalized inference tool (GIT) achieving the highest overall results among the models tested, a significant disparity exists in semantic reasoning capabilities, particularly with Macedonian. This gap highlights that cultural inference presents additional challenges in these contexts, underscoring the complexities of processing and understanding languages that lack sufficient resources and data support. Examples 5, 6, 7, and 8 provide more information on the performance of these models.

Image: A child wearing a firefighter costume holding a toy fire hose.

Example	Language / Question	CLIP	mPLUG	GIT
EX 5:	**English/**What job is the child pretending to do?	"Playing with a hose." **Analysis:** Misses the intent; does not identify the job role; overly literal.	"The child is pretending to be a firefighter." **Analysis:** Correct, though a bit mechanical; directly answers the question.	"The child is dressed as a firefighter and playing as if putting out a fire." **Analysis:** Correct and context-rich; provides a nuanced interpretation adding narrative context.
EX 6:	**Arabic/** ما الوظيفة التي يتظاهر الطفل أنه يؤديها؟ *Translation:* What job is the child pretending to do?	"الطفل يلعب بأنبوب مياه." ***Translation:*** "The child is playing with a water hose." **Analysis:** Lacks accuracy; describes the action only without identifying the job.	" الطفل يتظاهر بأنه رجل إطفاء." ***Translation:*** "The child is pretending to be a firefighter." **Analysis:** Correct but straightforward; lacks descriptive detail.	" الطفل يرتدي زي رجل إطفاء ويلعب وكأنه يطفئ حريقًا." ***Translation:*** "The child is wearing a firefighter's uniform and playing as if putting out a fire." **Analysis:** Accurate and rich in context; reflects cultural and narrative depth.
EX 7:	**Malay**/*Pekerjaan apa yang sedang dilakonkan oleh kanak-kanak itu*? *Translation:* What job is the child pretending to do?	"*Kanak-kanak bermain dengan hos air*." ***Translation:*** "The child is playing with a water hose." **Analysis:** Too literal; fails to identify the implied profession.	"*Kanak-kanak itu sedang berlakon sebagai seorang bomba*." ***Translation:*** "The child is acting as a firefighter." **Analysis:** Accurate but minimal; lacks expressive detail.	"*Kanak-kanak itu memakai pakaian bomba dan berlakon seperti sedang memadam kebakaran*." ***Translation:*** "The child is wearing a firefighter's outfit and acting as if putting out a fire." **Analysis:** Clear, descriptive, and contextually rich; includes visual and cultural cues.

EX 8:	**Macedonian /** Кое занимање се преправа детето дека го извршува? *Translation:* What job is the child pretending to do?	"Детето игра со црево за вода." ***Translation:*** "The child is playing with a water hose." **Analysis:** Inaccurate; focuses on the action, not the profession.	"Детето се преправа дека е пожарникар." ***Translation:*** "The child is pretending to be a firefighter." **Analysis:** Correct but lacks depth or narrative.	"Детето носи костим на пожарникар и се преправа дека гаси пожар." ***Translation:*** "The child is wearing a firefighter's costume and pretending to put out a fire." **Analysis:** Strong contextual and cultural understanding; accurately captures intent and detail.

The task of speech-to-image narrative generation tested the Whisper + GPT-4V pipeline. The study involved participants who provided spoken prompts in their native languages. These prompts were specifically tailored to describe a particular image scenario that was presented to them. The model utilized these diverse linguistic inputs to enhance its understanding and response generation about the described images, reflecting the nuances and contextual meanings embedded in each language. Generated visual and narrative outputs.

As illustrated in Table 3, the study's results reveal that Whisper's transcription accuracy was notably high across various languages, indicating its effectiveness in processing diverse linguistic inputs. However, in contrast, GPT-4V faced significant challenges in accurately representing cultural symbols within the generated images. This issue was particularly pronounced for the Macedonian and Malay cultures, where specific cultural nuances and representations were inadequately captured. Additionally, human evaluators observed a concerning trend of bias towards Western visual aesthetics in the output. This suggests that the model may not fully embrace or reflect the rich diversity of global artistic traditions. This discrepancy underscores the need for improvement in cultural representation to ensure more inclusive and accurate visual representations across diverse cultures and languages. Examples 9, 10, 11, and 12 provide insights into these points.

Table 3. Speech-to-image task – human evaluation scores (1-5)

Language	Fluency	Semantic alignment	Cultural relevance
English	4.6	4.5	4.4
Arabic	4.3	4.1	4.0
Malay	4.2	4.0	3.8
Macedonian	3.8	3.5	3.3

Example	Language/Spoken prompt	GPT-4v image output description	Analysis of sore
EX 9:	**Spoken prompt (in English):** A young girl is reading a book under a tree in a public park on a sunny afternoon.	The image shows a young girl sitting under a leafy tree, reading a colorful book. Sunlight filters through the leaves, and a park bench is nearby with flowers in bloom.	**High score:** **Fluency**: Whisper transcription is nearly perfect. **Semantic Alignment**: Visual scene matches narrative elements. **Cultural Relevance**: Matches typical Western imagery of public parks and childhood learning.
EX 10:	**Spoken Prompt (in Arabic):** *"طفل يرتدي جلابية يلعب أمام المسجد في الصباح الباكر."* *(A child wearing a jalabiya is playing in front of a mosque in the early morning.)*	A child stands in a courtyard with domed architecture in the background. The outfit resembles a robe, but the building lacks regional design cues. The mosque is rendered generically, resembling a cathedral dome.	**High, but lower score than English:** **Fluency:** Whisper handled Arabic well. **Semantic Alignment**: Scene broadly correct but lacks architectural specificity. **Cultural Relevance:** Mosque misrepresented-minarets missing, jalabiya not clear.
EX 11:	**Spoken Prompt (in Malay):** *"Seorang nelayan tua sedang menjala ikan di tepi pantai waktu subuh."* *(An old fisherman is casting his net by the beach at dawn.)*	An elderly man stands near water, casting a net. Scene is misty, lighting suggests morning, but setting looks like a lake in Europe. Clothing does not resemble Malay fishing attire.	**Average score:** **Fluency**: Whisper transcription was mostly accurate. **Semantic Alignment**: Correct action (casting net), time (dawn). **Cultural Relevance**: Missed local context—attire and beach setting not Southeast Asian.
EX 12:	**Spoken Prompt (in Macedonian):** *"Стар човек седи на чардак и пие кафе додека гледа во планините."* *(An old man is sitting on a 'čardak' drinking coffee while looking at the mountains.)*	An elderly man on a wooden balcony drink from a mug. The background includes snowy peaks, but architecture resembles a Swiss chalet. The čardak (Balkan-style veranda) is **not correctly represented**.	**Low Score** **Fluency**: Whispered slightly misrecognized "чардак." **Semantic Alignment**: General idea captured. **Cultural Relevance**: The visual elements were Westernized, losing their Balkan ambiance.

Translating cultural idioms and interpreting multimodal expressions required a comprehensive approach. This process involved presenting idioms or gestures that are deeply embedded in their respective cultures as pairs of images and text. Each pair was designed to showcase the idiom or gesture and contextualize its meaning within the culture it originates. Subsequently, models were tasked with translating these culturally rooted expressions or interpreting their meanings in another language, ensuring that nuances and cultural significance were preserved as much as possible. Such an endeavor highlights the complexity of language and culture, demonstrating the intricate connections between visual representations and verbal expressions.

As Table 4 shows, only GIT achieved an accuracy rate exceeding 60% across idiom translation tasks. Macedonian idioms were often misinterpreted, particularly those involving proverbs with visual metaphors. Qualitative reviewers noted that models lacked pragmatic understanding and contextual background for culturally embedded meaning.

Table 4. Correct interpretation rates (%) of cultural idioms

Model	Arabic → English	Malay → English	Macedonian → English
CLIP	48.5	46.0	38.0
mPLUG	60.2	58.1	49.5
GIT	66.3	61.0	54.3

Example	Language/Idiom/proverb	CLIP	mPLUG	GET
EX 13:	**Arabic**: بلغ السيل الزّبى *Literal meaning:* "The flood has reached the highest point." *Figurative meaning: Enough is enough / The last straw.*	The water is high. **Analysis**: Literal and misses the idiomatic meaning.	Things have gone too far." **Analysis**: Approximates the intended figurative meaning.	This situation has become unbearable. It is the last straw. **Analysis**: Correct figurative translation with expressive accuracy.
EX 14:	**Malay/** "Bagai melepaskan batuk di tangga" *Literal meaning:* "Like coughing on the stairs." *Figurative meaning: Doing something half-heartedly / Poor effort.*	A person coughing on stairs. **Analysis**: Literal; completely misses the figurative intent.	Doing something without seriousness. **Analysis**: Decent approximation.	Acting with no real effort or commitment, just for show. **Analysis**: Most accurate in context and tone.
EX 15:	**Macedonian**/"Да ти се крене косата" *Literal meaning:* "For your hair to stand up." *Figurative meaning: To be extremely scared or shocked.*	*Someone with their hair standing up.* **Analysis**: Interpreted literally, possibly as static electricity or hairstyle.	*To be frightened or alarmed.* **Analysis**: Close to correct meaning.	*A terrifying experience that makes your hair stand on end.* **Analysis**: Accurate and idiomatic; captures emotional resonance.

An analysis of over 200 erroneous or biased outputs revealed three key patterns: first, there is a significant tendency towards over-westernization, as visual outputs for non-Western languages often default to Eurocentric dress, architecture, or objects. Second, gender stereotyping is prevalent; for instance, descriptions of professional roles, such as doctor or teacher, frequently exhibit gender bias, with nurses often portrayed as female and engineers typically depicted as male. Lastly, there is a frequent occurrence of literal translation of figurative language, where idioms like "break one is back" are translated literally, thus missing their intended figurative meanings. Examples 13, 14, and 15 provide insights into these points.

As shown in Table 5, the results demonstrate that GIT has consistently performed the best across various multilingual and multimodal tasks, particularly in structured tasks. These include image captioning, which can generate detailed descriptions of images, and VQA, where it accurately responds to questions based on visual content. This consistency indicates a robust capability in understanding and processing different forms of data, making it a versatile tool in various application areas. In contrast, the combination of Whisper and GPT-4V has demonstrated impressive fluency in speech-to-image storytelling, enabling a seamless transition from verbal narratives to visual representations. However, despite its linguistic fluidity and creativity strengths, this model exhibits notable shortcomings regarding deep cultural grounding. This lack of artistic depth may hinder its effectiveness in conveying nuanced messages in diverse cultural contexts, vital for effective communication.

Table 5. Model capability comparison

Model	Multilingual text	Visual understanding	Speech input	Cultural sensitivity
CLIP	Moderate	Strong	No	Low
mPLU	Strong	Strong	Limited	Moderate
GIT	Very strong	Very strong	No	Moderate
Whisper + GPT-4V	Strong	Very strong	Strong	Moderate

It is essential to note that all the models assessed demonstrated significant performance gaps when handling low-resource languages. This issue is especially pronounced in tasks that require contextual understanding or figurative interpretation, such as idioms or cultural references that do not translate directly across languages. These gaps underscore the challenges of developing AI models that can perform equally well across all languages, particularly those underrepresented in training datasets. Human evaluators have raised significant concerns regarding the reliance of current models on Western-centric training datasets. This dependency often leads to cultural misrepresentation or semantic drift when these models are applied in non-Western contexts. Such issues can result in misunderstandings or misinterpretations, underscoring the need for more inclusive and representative training data to enhance the cultural competency of AI models. Addressing these gaps is crucial for the responsible and effective deployment of AI technologies in a globally diverse environment.

5. DISCUSSION

The findings of this study provide compelling evidence that MMMLMs hold substantial promise for advancing global communication. Nevertheless, they face persistent limitations, particularly in cultural contextualization and support for low-resource languages. While recent models such as GIT and mPLUG have achieved strong performance across structured tasks like image captioning and VQA, their effectiveness drops considerably when dealing with culturally nuanced content, idiomatic expressions, and languages with limited training data.

One of the most striking patterns in the results was the performance disparity between high-resource and low-resource languages. Models performed strongly in English and moderately well in Arabic and Malay, but showed apparent limitations in Macedonian. This reinforces findings from prior studies [49]-[51], which emphasize that language representation in training corpora directly correlates with downstream performance.

Although models like GIT demonstrated robust multilingual capabilities, the accuracy and semantic alignment in VQA and captioning tasks declined when applied to culturally unfamiliar domains. The tendency to overgeneralize or default to Western imagery, particularly in Malay speech-to-image outputs, reveals a bias in the composition of training data and pretraining objectives, which often center on English and European cultural contexts.

While MMMLMs can technically process multiple languages and modalities, their ability to reason across cultures and interpret context-specific meaning remains shallow. This was especially evident in the idiom translation tasks and visual representations of culturally significant scenes. Even when fluency and surface-level correctness were achieved, the pragmatic communication layer was frequently missed, leading to misinterpretations or culturally insensitive outputs. These failures align with the critique that current language models often lack "situated knowledge" [52] and instead operate from a universalist, decontextualized framework. The inability to grasp figurative language and symbolic meaning in underrepresented cultures suggests a need to fine-tune with culturally diverse datasets and to incorporate community-informed annotation strategies.

Integrating modalities-such as speech, text, and images-broadens the communicative potential of AI systems, enabling more naturalistic human-machine interaction**.** For example, the Whisper + GPT-4V pipeline demonstrated strong fluency and narrative coherence in multiple languages, confirming the viability of multimodal approaches for tasks such as education, digital storytelling, and translation. However, the cross-modal alignment remains brittle. A correct transcription of a spoken prompt does not guarantee a culturally accurate or semantically faithful image. This suggests that modalities are not yet sufficiently grounded in one another, and that alignment techniques (e.g., contrastive learning or cross-modal attention) must evolve to reflect syntactic correlations and sociocultural intent.

The recurring biases and failures documented in the study raise significant ethical concerns. Gender stereotyping in visual outputs, western-centric iconography, and the flattening of cultural diversity reflect deeper algorithmic hegemonies embedded in the training data and the architectures [53], [54]. As MMMLMs are increasingly deployed in cross-cultural contexts, ranging from humanitarian work to media production, such biases may not only miscommunicate but also actively reinforce harmful stereotypes. The study shows that human-in-the-loop evaluations remain essential. Human reviewers' ability to detect misinterpretation and assess cultural appropriateness offers a valuable layer of accountability that automated metrics cannot replicate.

Given the limitations observed, it is clear that future developments in MMMLMs must emphasize localization**,** linguistic diversity, and contextual adaptability. Expanding multimodal datasets to include lesser-known languages, such as different dialects and creoles, is crucial for promoting inclusivity in language processing. Additionally, integrating aspects of cultural studies, anthropology, and linguistics into creating and annotating these datasets is vital. Utilizing participatory research models enables speakers of marginalized languages to collaborate in designing evaluation tasks and interpreting model outputs, thus ensuring that the research reflects their experiences and viewpoints.

"Cultural fidelity" should become a key performance metric in evaluating multilingual AI. Furthermore, federated learning or modular architectures may allow for community-specific model fine-tuning without compromising generalizability. Expanding the proposed "cultural fidelity" metric requires transforming it from a purely qualitative construct into a measurable, multimodal benchmark that integrates human evaluation with computational modeling. Beyond subjective scoring by cultural experts, cultural fidelity can be operationalized through hybrid evaluation pipelines combining linguistic, visual, and contextual alignment metrics. For text outputs, this involves developing lexicons and semantic-similarity models trained on culturally localized corpora to assess whether the generated language reflects the idiomatic, pragmatic, and sociosemantic norms of the target culture. In visual outputs, embedding-based comparison tools-such as CLIP-like cross-modal alignment networks fine-tuned on region-specific datasets-can detect the presence or absence of culturally appropriate symbols, attire, architecture, or environmental contexts. Additionally, sentiment and discourse analysis models can be adapted to detect tone and politeness strategies typical of specific linguistic communities, providing quantifiable indicators of cultural resonance. Temporal and spatial metadata (e.g., festival timing, traditional settings) can further support automated checks for cultural realism in multimodal scenes. Importantly, a multilayered scoring framework should integrate these computational indicators with periodic human audits by native speakers and cultural practitioners, creating a feedback loop for continuous refinement. In this way, cultural fidelity becomes a dynamic composite metric-rooted in both algorithmic detection and community-informed validation—allowing MMMLMs to be evaluated not only for accuracy but for their capacity to represent and respect the diversity of human cultural expression.

The findings of this study intersect directly with broader debates in AI ethics, particularly those concerning decolonial AI and situated knowledge. The observed cultural biases, Western-centric visual outputs, and inconsistent performance across low-resource languages illustrate how current MMMLMs often reproduce epistemic hierarchies embedded in their training data. From a decolonial perspective, these models reflect the dominance of Euro-American data sources, linguistic norms, and aesthetic conventions-an issue that scholars argue perpetuates digital colonialism by marginalizing non-Western epistemologies and communicative practices. The disparities in Macedonian and Malay cultural representation, for instance, exemplify how algorithmic design and dataset composition privilege the perspectives of those with greater computational and linguistic visibility. Similarly, drawing from Haraway's notion of situated knowledge, the study reinforces that AI models cannot be "view from nowhere" systems; their outputs are shaped by specific historical, cultural, and material conditions. Acknowledging this situatedness necessitates integrating diverse cultural agents, community data contributors, and local evaluators into model development pipelines. Thus, the study's call for cultural fidelity and human-in-the-loop evaluation aligns with the ethical imperative to design AI that is context-aware, pluralistic, and responsive to marginalized voices, challenging the myth of universal intelligence in favor of socially grounded, culturally accountable machine reasoning.

6. CONCLSUION

This study explored the capacities and limitations of MMMLMs-specifically, GIT, mPLUG, CLIP, and Whisper + GPT-4V-in fostering equitable global communication across diverse linguistic and cultural contexts. The results reveal substantial progress in multimodal reasoning and multilingual generalization, yet persistent disparities remain, particularly in the artistic and semantic grounding of low-resource languages. At the same time, high-resource languages such as English and Arabic benefited from richer representation in training corpora, languages like Macedonian, though Indo-European, performed worse due to their limited digital presence. These findings underscore that linguistic inequity in AI arises not from inherent linguistic complexity but from systemic data and resource imbalances entrenched in current model development pipelines.

Beyond technical insights, this study advances a broader ethical argument: the next generation of MMMLMs must not only be accurate but also culturally accountable. The inclusion of culturally informed evaluation protocols and the introduction of the "cultural fidelity" metric provide a pathway toward assessing models not just by computational benchmarks but by their respect for cultural nuance, symbolism, and pragmatic meaning. These frameworks emphasize that communicative equity requires valuing diverse epistemologies and linguistic traditions as integral to model design, rather than treating them as peripheral testing conditions. The findings align with critical debates in AI ethics and decolonial AI, reinforcing the notion that technological neutrality is a myth-models inevitably reflect the social, historical, and geographic biases of their data sources.

The policy implications are profound. Governments, funding agencies, and research consortia must prioritize targeted investment in low-resource data curation to ensure the sustainable development of annotated multimodal corpora for marginalized languages. Such initiatives should adopt community-based and participatory approaches, empowering local speakers and cultural practitioners to guide data annotation,

contextual validation, and interpretation. Funding bodies should incentivize the creation of open-access, ethically sourced datasets and require cultural diversity audits for publicly funded AI projects. At the same time, international policy frameworks-such as those developed by UNESCO, the European Union, and the African Union-should promote collaborative infrastructure for data sharing, multilingual benchmarking, and cross-regional capacity-building in AI research.

Furthermore, academic institutions and AI developers should embed linguistic justice, transparency, and algorithmic accountability into their design principles. Mandating that commercial and research models disclose training data sources, language distributions, and evaluation protocols would support greater fairness and reproducibility. Ethical guidelines must evolve from voluntary principles to regulatory standards, ensuring that cultural representation and linguistic inclusion are measurable and enforceable aspects of AI governance.

Ultimately, achieving equitable multimodal-multilingual AI demands more than technical sophistication-it requires a shift in global priorities. By coupling innovation with inclusion, data with diversity, and progress with policy, MMMLMs can become tools not of cultural homogenization but of mutual understanding. Investing in low-resource language infrastructure, culturally grounded evaluation, and interdisciplinary collaboration is therefore not only an ethical responsibility but a strategic imperative for building AI systems that genuinely reflect and serve the world's linguistic and cultural plurality.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohammad Awad AlAfnan	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	
Siti Fatimah MohdZuki		✓		✓	✓	✓		✓	✓	✓	✓			
Shefa Mohammad AlAfnan	✓		✓				✓	✓		✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [initials, MAA]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] V. Voronkova, G. Vasyl'chuk, V. Nikitenko, Y. Kaganov, and N. Metelenko, *Transformation of digital education in the era of the fourth industrial revolution and globalization*. Izdevnieciba "Baltija Publishing," 2023.

[2] Z. Chen *et al.*, "Evolution and prospects of foundation models: from large language models to large multimodal models," *Computers, Materials & Continua*, vol. 80, no. 2, pp. 1753–1808, 2024, doi: 10.32604/cmc.2024.052618.

[3] V. Nataraj *et al.*, "Generative AI in multimodal cross-lingual dialogue system for inclusive communication support," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, Aug. 2024, pp. 204–209, doi: 10.1109/IRI62200.2024.00051.

[4] J. Y. Koh, D. Fried, and R. Salakhutdinov, "Generating images with multimodal language models," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[5] P. Selvakumar and T. C. Manjunath, "AI in text paraphrasing," in *Using AI Tools in Text Analysis, Simplification, Classification, and Synthesis*, 2025, pp. 349–374.

[6] R. Shahmerdanova, "The role of translation in global diplomacy and international relations," *Journal of Azerbaijan Language and Education Studies*, vol. 2, no. 1, pp. 34–48, Jan. 2025, doi: 10.69760/jales.2025001003.

[7] Y. M. Al-Worafi, "Technology in public health education in developing countries," in *Handbook of Medical and Health Sciences in Developing Countries*, Cham: Springer International Publishing, 2024, pp. 1–20.

[8] B. Emihovich, "Implementing universal design for learning in online courses to support multilingual students in higher education," *The CATESOL Journal*, vol. 35, no. 1, Sep. 2024, doi: 10.5070/B5.34833.

[9] E. Dritsas, M. Trigka, C. Troussas, and P. Mylonas, "Multimodal interaction, interfaces, and communication: a survey," *Multimodal Technologies and Interaction*, vol. 9, no. 1, p. 6, Jan. 2025, doi: 10.3390/mti9010006.

[10] J. Jiang *et al.*, "A review of transformer models in drug discovery and beyond," *Journal of Pharmaceutical Analysis*, vol. 15, no. 6, p. 101081, Jun. 2025, doi: 10.1016/j.jpha.2024.101081.

[11] D. Gromann, "Neural language models for the multilingual, transcultural, and multimodal Semantic Web," *Semantic Web*, vol. 11, no. 1, pp. 29–39, Jan. 2020, doi: 10.3233/SW-190373.

[12] J. Armitage *et al.*, "MLM: A benchmark dataset for multitask learning with multiple languages and modalities," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, Oct. 2020, pp. 2967–2974, doi: 10.1145/3340531.3412783.

[13] L. Tay, S. E. Woo, L. Hickman, B. M. Booth, and S. D'Mello, "A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in Psychological Assessment," *Advances in Methods and Practices in Psychological Science*, vol. 5, no. 1, Jan. 2022, doi: 10.1177/25152459211061337.

[14] M. Balaban, L. Hamann, G. Khais, A. A. Saad, A. Maraee, and A. Sturm, "Mediation-based MLM in USE," in *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, Sep. 2024, pp. 818–827, doi: 10.1145/3652620.3688214.

[15] A. Stojanov, I. Toskov, T. Rompf, and M. Püschel, "SIMD intrinsics on managed language runtimes," in *Proceedings of the 2018 International Symposium on Code Generation and Optimization - CGO 2018*, 2018, pp. 2–15, doi: 10.1145/3179541.3168810.

[16] W. Zhang and J. Xu, "Graded subsidy policy-based equilibrium strategy applied to investment in electric vehicle chargers: a case study in Chengdu," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2678, no. 3, pp. 698–714, Mar. 2024, doi: 10.1177/03611981231182713.

[17] S. Mu, L. Han, and Z. (Edward) Wen, "Language portraits going digital and multimodal: deciphering the translanguaging space and linguistic repertoires among multilinguals," *Language and Education*, vol. 39, no. 1, pp. 132–153, Jan. 2025, doi: 10.1080/09500782.2024.2317946.

[18] A. Henlein *et al.*, "An outlook for AI innovation in multimodal communication research," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14711 LNCS, 2024, pp. 182-234.

[19] M. A. AlAfnan, "Large language models as computational linguistics tools: a comparative analysis of ChatGPT and Google machine translations," *Journal of Artificial Intelligence and Technology*, vol. 5, pp. 20–32, Jun. 2024, doi: 10.37965/jait.2024.0549.

[20] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: a survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1-40, Feb. 2024, doi: 10.1145/3605943.

[21] W. Khan, A. Daud, K. Khan, S. Muhammad, and R. Haq, "Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends," *Natural Language Processing Journal*, vol. 4, p. 100026, Sep. 2023, doi: 10.1016/j.nlp.2023.100026.

[22] M. V. Koroteev, "BERT: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021, [Online]. Available: http://arxiv.org/abs/2103.11943.

[23] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 55–65, doi: 10.18653/v1/D19-1006.

[24] L. Floridi and M. Chiriatti, "GPT-3: its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[25] Y. Ma, "Cross-language text generation using mBERT and XLM-R: English-Chinese translation task," in *2024 International Conference on Machine Intelligence and Digital Applications*, May 2024, pp. 602–608, doi: 10.1145/3662739.3672320.

[26] M. H. et Al., "CLIP and complementary methods," *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 19, Mar. 2021, doi: 10.1038/s43586-021-00023-4.

[27] Kai-Qing Zhou and Hatem Nabus, "The ethical implications of DALL-E: opportunities and challenges," *Mesopotamian Journal of Computer Science*, vol. 2023, pp. 16–21, Feb. 2023, doi: 10.58496/MJCSC/2023/003.

[28] S. Wang, C.-H. Yang, J. Wu, and C. Zhang, "Can whisper perform speech-based in-context learning?," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 13421-13425, doi: 10.1109/ICASSP48485.2024.10446502.

[29] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: a universal encoder for vision and language by cross-modal pre-training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11336–11344, Apr. 2020, doi: 10.1609/aaai.v34i07.6795.

[30] M. Ni *et al.*, "M 3 P: learning universal representations via multitask multilingual multimodal pre-training," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 3976–3985, doi: 10.1109/CVPR46437.2021.00397.

[31] C. Li *et al.*, "mPLUG: effective and efficient vision-language learning by cross-modal skip-connections," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 7241–7259, doi: 10.18653/v1/2022.emnlp-main.488.
[32] S. Kanthed, "From code to cloud: the role of GitOps, GitHub, and GitLab in modern DevOps," *Journal of Technological Innovations Est*, 2025, [Online]. Available: http://jtipublishing.com/jti.
[33] M. Awad, "DeepSeek vs. ChatGPT: a comparative evaluation of AI tools in composition, business writing, and communication tasks," *Journal of Artificial Intelligence and Technology*, vol. 5, pp. 202–210, 2025, [Online]. Available: https://ojs.istp-press.com/jait/article/view/740/563.
[34] A. Jain *et al.*, "MURAL: multimodal, multitask representations across languages," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 3449–3463, doi: 10.18653/v1/2021.findings-emnlp.293.
[35] J. Chen, Z. Zeng, Y. Lin, W. Li, Z. Ma, and M. Shou, "LiveCC: Learning Video LLM with streaming speech transcription at scale," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29083–29095.
[36] M. A. AlAfnan, "Who is feeding whom? A linguistic inquiry into Human–AI relationship," *Journal of Artificial Intelligence and Technology*, Nov. 2025, doi: 10.37965/jait.2025.0789.
[37] L. Yang, Y. Zhang, S. Kang, Z. Wang, and C. Wu, "Microplastics in soil: a review on methods, occurrence, sources, and potential risk," *Science of The Total Environment*, vol. 780, p. 146546, Aug. 2021, doi: 10.1016/j.scitotenv.2021.146546.
[38] M. A. AlAfnan, "Taxonomy of educational objectives: teaching, learning, and assessing in the information and artificial intelligence era," *Journal of Curriculum and Teaching*, vol. 13, no. 4, p. 173, Aug. 2024, doi: 10.5430/jct.v13n4p173.
[39] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021, pp. 610–623, doi: 10.1145/3442188.3445922.
[40] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy, "SpanBERT: improving pre-training by representing and predicting spans," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 64-77, Dec. 2020, doi: 10.1162/tacl_a_00300.
[41] E. Pavlick, "Symbols and grounding in large language models," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 381, no. 2251, Jul. 2023, doi: 10.1098/rsta.2022.0041.
[42] D. H. Park *et al.*, "Multimodal explanations: justifying decisions and pointing to the evidence," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 8779–8788, doi: 10.1109/CVPR.2018.00915.
[43] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: multilingual English-German image descriptions," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74, doi: 10.18653/v1/W16-3210.
[44] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. P. Levy, "On the predictive power of neural language models for human real-time comprehension behavior," *Proceedings for the 42nd Annual Meeting of the Cognitive Science Society: Developing a Mind: Learning in Humans, Animals, and Machines, CogSci 2020*, pp. 1707–1713, 2020.
[45] T. Brummaier *et al.*, "Cohort profile: molecular signature in pregnancy (MSP): longitudinal high-frequency sampling to characterise cross-omic trajectories in pregnancy in a resource-constrained setting," *BMJ Open*, vol. 10, no. 10, p. e041631, Oct. 2020, doi: 10.1136/bmjopen-2020-041631.
[46] M. A. AlAfnan, "Artificial Intelligence and Language: bridging Arabic and English with technology," *Journal of Ecohumanism*, vol. 3, no. 8, Nov. 2024, doi: 10.62754/joe.v3i8.4961.
[47] A. Rahman and M. Wali Ur, "Optimizing large language models for edge devices: a comparative study on reputation analysis item type electronic thesis; text," Univ. of Arizona, 2023.
[48] UNESCO, "Key facts UNESCO's recommendation on the ethics of artificial intelligence," *United Nations Educational, Scientific and Cultural Organization*, 2023.
[49] L. K. Senel, B. Ebing, K. Baghirova, H. Schuetze, and G. Glavaš, "Kardeş-NLU: transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for turkic languages," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 1672–1688, doi: 10.18653/v1/2024.eacl-long.100.
[50] M. A. Hasan, P. Tarannum, K. Dey, I. Razzak, and U. Naseem, "Do large language models speak all languages equally? A comparative study in low-resource settings," *arXiv preprint arXiv:2408.02237*, 2024, [Online]. Available: http://arxiv.org/abs/2408.02237.
[51] V. Protasov, E. Stakovskii, E. Voloshina, T. Shavrina, and A. Panchenko, "Super donors and super recipients: studying cross-lingual transfer between high-resource and low-resource languages," in *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, 2024, pp. 94–108, doi: 10.18653/v1/2024.loresmt-1.10.
[52] D. Haraway, "Situated knowledges: the science question in feminism and the privilege of partial perspective," *Philosophical Literary Journal Logos*, vol. 32, no. 1, pp. 237–271, 2022, doi: 10.22394/0869-5377-2022-1-237-268.
[53] R. G. Bender *et al.*, "Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the global burden of disease study 2021," *The Lancet Infectious Diseases*, vol. 24, no. 9, pp. 974–1002, Sep. 2024, doi: 10.1016/S1473-3099(24)00176-2.
[54] M. A. Alafnan, "Technical Report Writing Efficiency Using AI-powered tools: opportunities, challenges, and future directions," *Journal of Artificial Intelligence and Technology*, vol. 5, pp. 270–277, 2025, doi: 10.37965/jait.2025.0729.

BIOGRAPHIES OF AUTHORS

Mohammad Awad AlAfnan is an associate professor of Applied Linguistics/Business Communication. He participated in a number of international conferences and published journal articles and books on AI in Education, Teaching and learning, critical discourse analysis, and mass/social media. He taught undergraduate and postgraduate courses, moderated master's theses and dissertations, and supervised Ph.D. dissertations. He can be contacted at email: Mohammad.al-afnan@aum.edu.kw.

Siti Fatimah MohdZuki studied Electrical Engineering at RMIT University in Melbourne, Australia, and holds a Master's degree in Computer Science from Universiti Teknologi MARA (UiTM) in Shah Alam. She has published extensively on topics related to Artificial Intelligence, as well as teaching and learning methodologies. She can be contacted at email: SitiFatimahMohdZUki@yahoo.com.

Shefa Mohammad AlAfnan is affiliated with the International Islamic University, Kuala Lumpur, Malaysia, and has diverse research interests, including learning and communication. Her publications include work on educational taxonomies and learning frameworks. She can be contacted at email: ShefaMohammadAlAfnan@yahoo.com.