MMLoSo 2025

# 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact

## Proceedings of the Workshop

December 23, 2025

The MMLoSo organizers gratefully acknowledge the support from the following sponsors.

## In Collaboration With

जनजातीय कार्य मंत्रालय
MINISTRY OF
TRIBAL AFFAIRS
सत्यमेव जयते

IFML

©2025 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
`acl@aclweb.org`

ISBN 979-8-89176-311-1

# Preface

Welcome to the First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025), co-located with IJCNLP-AACL 2025 in Mumbai, India.

This workshop brings together researchers at the intersection of multimodal learning, NLP, and AI for social good, with a focus on low-resource and underserved settings. We aim to bridge the gap between the growing capabilities of multimodal machine learning models and the urgent needs of real-world applications in under-resourced, marginalized, or data-constrained settings.

We received 26 submissions in total. After desk rejecting 2 papers and 1 withdrawal, 23 papers entered the peer review process. We accepted 14 papers (10 oral presentations and 4 poster presentations) after a rigorous review process, representing an acceptance rate of 53.8% (14 out of 26 submissions).

The workshop features contributions on learning with missing or incomplete modalities, few-shot and zero-shot learning in multimodal contexts, multilingual representation learning, ethical and interpretable AI, and applications in social good including ecological monitoring, public health, language documentation, and crisis response.

We thank all authors for their excellent contributions, our program committee for their thorough reviews, our keynote speakers for their insights, and our student volunteers for their invaluable assistance in making this workshop a success.

We hope this workshop fosters collaboration and innovation in developing robust and inclusive multimodal systems that can operate effectively under data constraints for social impact.

MMLoSo 2025 Workshop Organizers
Ankita Shukla, Sandeep Kumar, Amrit Singh Bedi, and Tanmoy Chakraborty

# Organizing Committee

**Workshop Organizers**

Ankita Shukla, University of Nevada, Reno
Sandeep Kumar, Indian Institute of Technology Delhi
Amrit Singh Bedi, University of Central Florida
Tanmoy Chakraborty, Indian Institute of Technology Delhi

**Student Volunteers**

Pooja Singh, Indian Institute of Technology Delhi
Debarchan Basu, Indian Institute of Technology Delhi
Vaibhav Sharma, Indian Institute of Technology Delhi
Harsh Chauhan, Indian Institute of Technology Delhi
Sanjeev Singh, Indian Institute of Technology Delhi
Shashwat Bhardwaj, Indian Institute of Technology Delhi

# Program Committee

## Program Chairs

Ankita Shukla, University of Nevada, Reno
Sandeep Kumar, Indian Institute of Technology Delhi
Amrit Singh Bedi, University of Central Florida
Tanmoy Chakraborty, Indian Institute of Technology Delhi

## Program Committee

Anil Sharma, Samsung R&D
Gullal S. Cheema, Leibniz University Hannover
Sinjini Mitra, Ground Truth Analytics
Jyotismita Barman, Indian Institute of Technology Delhi
Abhishek Gupta, Indian Institute of Technology Delhi
Mohit Kataria, Indian Institute of Technology Delhi
Shruti Kshirsagar, Wichita State University
Jisoo Lee, Arizona State University
Nikita Malik, Indian Institute of Technology Delhi
Gourab Panda, Indian Institute of Technology Delhi
Aditya Hemant Shahane, Indian Institute of Technology Delhi
Pooja Singh, Indian Institute of Technology Delhi
Vipul Kumar Singh, Indian Institute of Technology Delhi

# Table of Contents

*Artwork Interpretation with Vision Language Models: A Case Study on Emotions and Emotion Symbols*
Sebastian Padó and Kerstin Thomas . . . . . 1

*Toward Automatic Safe Driving Instruction: A Large-Scale Vision Language Model Approach*
Haruki Sakajo, Hiroshi Takato, Hiroshi Tsutsui, Komei Soda, Hidetaka Kamigaito and Taro Watanabe . . . . . 13

*TRepLiNa: Layer-wise CKA+REPINA Alignment Improves Low-Resource Machine Translation in Aya-23 8B*
Toshiki Nakai, Ravikiran Chikkala, Lena Oberkircher, Nicholas Jennings, Natalia Skachkova, Tatiana Anikina and Jesujoba Oluwadara Alabi . . . . . 25

*MemeGuard: Transformer-Based Fusion for Multimodal Propaganda Detection in Low-Resource Social Media Memes*
Md. Mohiuddin, Kawsar Ahmed, Shawly Ahsan and Mohammed Moshiul Hoque . . . . . 35

*Language as a Label: Zero-Shot Multimodal Classification of Everyday Postures under Data Scarcity*
Ming Ze Tang and Jubal Chandy Jacob . . . . . 48

*BengaliFig: A Low-Resource Challenge for Figurative and Culturally Grounded Reasoning in Bengali*
Abdullah Al Sefat . . . . . 58

*Domain-Specific Adaptation for ASR through Text-Only Fine-Tuning*
Betty Kurian, Abhinav Upadhyay and Abhijeet Sengupta . . . . . 78

*Towards Blind and Low-Vision Accessibility of Lightweight VLMs and Custom LLM-Evals*
Shruti Singh Baghel, Yash Pratap Singh Rathore, Anurag Pradhan, Sushovan Jena, Arnav Bhavsar, Amit Shukla and Pawan Goyal . . . . . 86

*Evaluating IndicTrans2 and ByT5 for English–Santali Machine Translation Using the Ol Chiki Script*
Kshetrimayum Boynao Singh, Asif Ekbal and Partha Pakray . . . . . 95

*Challenge Track: LoRAs in All Directions: Directional Adapters and Noisy-Channel Reranking for Indic MT*
Sajay Raj . . . . . 101

*Challenge Track: Breaking Language Barriers: Adapting NLLB-200 and mBART for Bhilli, Gondi, Mundari, and Santali Without Source Language Proficiency*
Paul Kamau . . . . . 106

*Challenge Track: Divide and Translate: Parameter Isolation with Encoder Freezing for Low-Resource Indic NMT*
Vaibhav Kanojia . . . . . 109

*Challenge Track: JHARNA-MT: A Copy-Augmented Hybrid of LoRA-Tuned NLLB and Lexical SMT with Minimum Bayes Risk Decoding for Low-Resource Indic Languages*
Dao Sy Duy Minh, Trung Kiet Huynh, Tran Chi Nguyen, Phu Quy Nguyen Lam, Phu-Hoa Pham, Nguyễn Đình Hà Dương, Dien Dinh and Long HB Nguyen . . . . . 114

*Findings of the MMLoSo 2025 Shared Task on Machine Translation into Tribal Languages*
Pooja Singh, Sandeep Chatterjee, Gullal S. Cheema, Amrit Singh Bedi, Tanmoy Chakraborty, Sandeep Kumar and Ankita Shukla . . . . . 121

# Artwork Interpretation with Vision Language Models: A Case Study on Emotions and Emotion Symbols

**Sebastian Padó*** **and** **Kerstin Thomas†**
* Institute for Natural Language Processing (IMS)
† Institute for Art History (IKG)
University of Stuttgart, Germany
`{sebastian.pado@ims,kerstin.thomas@ikg}.uni-stuttgart.de`

## Abstract

Emotions are a fundamental aspect of artistic expression. Due to their abstract nature, there is a broad spectrum of emotion realization in artworks. These are subject to historical change and their analysis requires expertise in art history. In this article, we investigate which aspects of emotional expression can be detected by current (2025) vision language models (VLMs). We present a case study of three VLMs (Llava-Llama and two Qwen models) in which we ask these models four sets of questions of increasing complexity about artworks (general content, emotional content, expression of emotions, and emotion symbols) and carry out a qualitative expert evaluation. We find that the VLMs recognize the content of the images surprisingly well and often also which emotions they depict and how they are expressed. The models perform best for concrete images but fail for highly abstract or highly symbolic images. Reliable recognition of symbols remains fundamentally difficult. Furthermore, the models continue to exhibit the well-known LLM weakness of providing inconsistent answers to related questions.

## 1 Introduction

Emotions are a privileged aspect of artistic expression in the visual arts (Tan, 2000). Arguably, many artworks actually address the emotions of the viewer far more directly than the intellect, in order to gain privileged access to viewer's morals, beliefs, values, and worldview. In fact, for a long time, the recommendations for the art of persuasive speech (rhetoric) and those for the visual arts have been directly related (Barthes, 1977).

In recent years, digital access to artwork has scaled up considerably and has become an important branch of (digital) cultural heritage curation (Näslund and Wasielewski, 2020). For example, Prometheus (Dieckmann, 2010) is a distributed archive which as of now (August 2025) provides a unified interface to access almost 4 million images from art, culture, and history. Such archives offer the opportunity for 'scalable reading' (Weitin, 2017) – or, in this case, more accurately 'scalable viewing' – studies in which automated analysis on large data sets is combined with a focused manual analysis on smaller samples and which have become widely used in other areas of Digital Humanities, notably literary studies (de Sá Pereira, 2019). This approach would be particularly attractive for research questions that inherently involve large number of images, such as: Which elements in images have constant significance for emotional expression, and which exhibit variance? How have such elements changed historically? What cultural differences are there in emotional expression?

In the textual modality – also prominent in cultural heritage – the analysis of emotions has taken major steps in the last ten years. Neural approaches can detect emotions significantly more accurately than previous approaches (Nandwani and Verma, 2021). They no longer rely on dictionaries of keywords but can recognize emotional contexts, including stylistic devices such as irony, while also taking global document meaning into account. Such models have also been used successfully for the analysis of emotional content in literary texts (Kim and Klinger, 2019). Recently, vision language models (VLMs) generalize the success of text-based methods by tightly integrating language and image information, enabling cross-modal information transfer: They enable the generation of images from textual descriptions, conversely the generation of textual descriptions for images, and the answering of textual questions on images (Antol et al., 2015).

This development has the potential to carry out emotion analysis on artwork with VLMs. However, the success of this approach is all but guaranteed. Due to their abstract nature, emotions are realized visually in a wide variety of ways. Artists can draw on the study of natural emotional expression,

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 1–12
December 23, 2025 ©2025 Association for Computational Linguistics

on facial expressions and gestures. Alternatively, an emotional expressions can result from various abstract elements, such as colors, shapes, compositional schemes, or symbols: signs with a secondary conventionalized level of meaning (Cassirer, 1923). Since their meaning can only be analyzed in context and is subject to historical change, the analysis of emotional content of artworks is generally assumed to require art historical expertise. Additionally, on the technical level, it is well known that VLMs, despite a surprisingly good understanding of individual aspects, struggle to develop a globally coherent understanding and can suffer from hallucinations (Huang et al., 2025). Substantive studies on (properties of) artworks however have to assume that the automatic analysis is largely accurate, or that remaining errors are at least distributed as randomly as possible.

To our knowledge, there are no studies that gauge the concrete quality of emotion-related analyses that fall out of current VLMs. Our study aims to fill this gap. We select 38 images of artworks (mostly of paintings, but also including sculptures and photographs), present these images to three current vision-language models and ask them a catalog of eight questions, ranging from a pure description of the content to the interpretation of the emotions and any symbols used. We qualitatively evaluate the models' outputs by hand to understand how differentiated the models generally perceive the different aspects of our images.

We obtain mixed results: VLMs are capable of correctly recognizing image content, often including emotions and the artistic means used to express them. However, recognition generally seems to be based on conventionalized patterns and fails with novel combinations. Complex and symbolic images also pose difficulties. Our conclusion is that current VLMs are already sufficiently accurate for some scalable reading research questions and for some types of images, but not yet for others.

## 2 Background and Related Work

### 2.1 Emotions in Art History

From an art historical perspective, emotions are central to the artistic impact. A work of art is meant to persuade, move, and stimulate thought. However, how works of art achieve this effect is often far from obvious.

One possibility is for them to use means of expression from natural life that are based on anthropological patterns: an open mouth with raised corners indicates laughter, while narrowed eyes and drooping corners indicate sadness. The same applies to physical means of expression such as posture and gestures. The expressive values are modular: one expressive value is reinforced by another. They are also quite constant over time within a cultural area. For this reason, they are referred to in psychology as 'basic emotions' (Ekman, 1999).

As an alternative to means of expression based on gestures and facial expressions, artists employ less clearly legible means of expression in their works, such as certain colors, shapes, compositional devices, or symbols – e.g., heart for love. These act alongside the anthropological means (such as smiles), reinforcing them and leading to specific emotional expressions. The knowledge needed to understand these means is culturally and historically specific, and reading them correctly requires a historical understanding. For this reason, many researchers call for a model of 'historical emotion research' in addition to the more universal model of basic emotions (Stearns and Stearns, 1985; Rosenwein, 2010; Frevert et al., 2011; Matt, 2011; Plamper, 2012).

Art historical research rarely distinguishes these two scenarios. Instead, emotions in works of art are attributed predominantly on the basis of concrete, codified expression schemes and symbols. The more abstract means of conveying emotions are rarely analyzed in detail, but rather take a back seat in the shape of general and diffuse descriptions of impressions. Our study takes steps towards remedying this situation, using VLMs as a pre-theoretical device that gives – at least in theory – equal importance to the different elements in the image. In practice, of course, this depends on the materials on which the models we use were pretrained.

### 2.2 Language and Vision-Language Models

Attempts to support the analysis of artworks with automatic methods can at this point build on the technological progress in AI/NLP of the last ten year, notably Language Models (LM) based on transformers (Vaswani et al., 2017). In contrast to earlier approaches, transformer-based LMs can consider an extensive linguistic context and can be (pre-)trained on large datasets since their training parallelizes well. From a user perspective, an important development in recent years is the emergence of instruction-tuned models (Brown et al., 2020) that are able to answer textual questions without task-specific training, which makes them directly applicable for interactive text-based querying by domain experts.

Vision-Language Models (VLMs) go beyond language and strive to develop a shared understanding of information from multiple modalities – typically text and visual data. Conceptually, this is achieved by bringing together embeddings from both modalities. Most VLMs consist of an encoder for the image, whose output is projected onto the embedding space of a language model that represents text meaning. Early models such as VilBERT (Lu et al., 2019) still use conventional neural networks such as CNNs for the visual encoders. Current models achieve a new level of quality by using contrastively trained image encoders such as CLIP (Radford et al., 2021), trained to match very large sets of image-description pairs. The resulting models capture semantic concepts across domains. VLMs inherit the ability to accept textual instructions and questions and generate verbal output from their component LMs, enabling them to perform ’visual question answering’ (Antol et al., 2015).

However, current models are clearly not perfect. Like LMs, VLMs exhibit a tendency toward hallucinations (Liu et al., 2024), which can be described as a cross-modal inconsistency between image and description. Judging from experience with LMs, we might expect that VLMs will have more difficulty the less experience they have with a type of image, a type of question, or both.

### 2.3 Artwork Analysis with Language-Vision Models

To our knowledge, there are only two previous studies in NLP that study the outputs of VLMs for artworks. Hayashi et al. (2024) create a corpus pairing artworks with their corresponding Wikipedia articles. They ask VLMs to generate texts corresponding to article parts (sections, subsections, etc.) and evaluate the output against the actual article parts with natural language generation metrics, both in terms of textual overlap (e.g., BLUE and ROUGE) and in terms of entity match metrics. Ozaki et al. (2025) extend this paradigm to multiple languages and in addition assess the capability of tuning to improve the models’ outputs. These studies found somewhat mixed results: The VLMs generally used a certain amount of ground-truth entities, and LoRA tuning improved results, but this was only true for English, and performance for other languages was lower throughout.

## 3 Experimental Setup

### 3.1 Motivation

In our study, we use the same fundamental generation paradigm as Ozaki et al. (2025) – prompting VLMs with a pair of picture and question – but combine it with a different evaluation paradigm. We do not evaluate the generated texts against a ground truth, for two reasons: (a) NLG metrics, such as entity coverage, are difficult to translate into insights, and it is hard to understand from the previous studies how well the VLMs really understand the images; (b) the questions are determined by the Wikipedia article structures and thus include questions that are not visual in nature (’What is the history of the Mona Lisa?’) while they do not include topic-specific questions – such as those concerning our topic of interest, emotions.

To address concern (a), to carry out a qualitative analysis, presenting the VLM outputs to two experts (the authors) to annotate for reasonableness.[1] We thus trade a fully evaluation procedure against a (hopefully) more detailed understanding of the capabilities and limits of the VLMs.

To realize this potential, we address concern (b) by defining a set of eight questions that we ask the VLMs for each image, listed in Table 1. We formulate the questions in English, since current LLMs and VLMs are trained predominantly with English data (Zhang et al., 2024). Therefore, model errors observed for English prompts can be more reliably interpreted as deficits in conceptual understanding than linguistic shortcomings (Ozaki et al., 2025; Qin et al., 2025).

The questions progress from basic descriptions of form and content (Q 1 and 2) to the interpretation of the displayed emotions (Q 3 to 5) and more nuanced characteristics of these emotions, such as the means used for representation (Q 6), the use of symbols (Q 7), and their intensity (Q 8). The recurring request for brevity was necessary to neutralize the models’ well-known tendency to give lengthy answers (Wang and Zhou, 2024).

We also note that our case study is focused on understanding the visual understanding of the models in their default as-published state. For this reason, we do not provide the images’ titles to the VLMs (Hayashi et al., 2024) – so that all information has to be inferred from visual information – and we do

---

[1] The quality assessments largely correspond, so the results we report below represent a consensus between the authors.

| | |
|---|---|
| Q 1 | Is this a painting, a drawing, a sculpture, or something else? Be brief. |
| Q 2 | What does this artwork show? Be brief. |
| Q 3 | Does this artwork involve an emotion? Give a yes/no answer and a brief justification. |
| Q 4 | Is the emotion shown by this artwork a positive or a negative one? Give a one-word answer (positive/negative) and a brief justification. |
| Q 5 | What emotion is shown by this artwork? Be brief. |
| Q 6 | How is the emotion shown by this artwork expressed artistically? Do not write more than a short paragraph. |
| Q 7 | Does this artwork use a symbol to visualize emotion, and if it does, how? Do not write more than a short paragraph. |
| Q 8 | How intense is the emotion depicted? Be brief. |

Table 1: Questions posed to VLMS for image interpretation: Basic description (Q 1–2), Emotion recognition (Q 3–5), Emotion expression (Q 6–8)

not experiment with fine-tuning the VLMs (Ozaki et al., 2025).

### 3.2 Selection of VLMs

The largest and best-performing LLMs and VLMs are all proprietary and can only be used via company-controlled APIs, which is problematic from the perspective of transparency and reproducibility (Liesenfeld et al., 2023). We therefore limit our study to three comparatively small VLMs that have open weights, i.e., can be downloaded in their entirety, and are small enough to be executed locally on a single 48GB GPU core.

Our first model is LLAVA-LLAMA-8B. It is a member of the Llava 'Large Language and Vision Assistant' (Liu et al., 2023) model family. It combines CLIP as an image encoder with Meta AI's large multilingual Llama 8B LM. The two other models are variants of the Alibaba Qwen-VL architecture (Bai et al., 2025) which combines a CLIP image encoder with the Qwen 2.5 multilingual LMs. We use a version with 7B parameters (QWEN-7B) and one with 32B parameters quantized with AWQ (QWEN-32B-AWQ, Lin et al. (2024)).

The three models are broadly similar in that they use similarly structured image encoders as well as similar transformer-based LMs. However, they behave significantly differently in practice, which is due to the fact that they were trained by different developers on different data. Unfortunately, we do not have sufficiently detailed information about the training process of any of the models to predict specific behavioral patterns.

### 3.3 Selection of Images

We compile a set of 38 images of artworks from the Prometheus image archive (Dieckmann, 2010), a huge data base of (among other domains) artworks from over 100 different organizations. We select artworks from a wide range of periods (including the Middle Ages, Renaissance, Baroque, and the 18th to 20th centuries) that have relatively clear emotional content. Our sample comprises mostly paintings, supplemented by some drawings, photographs, and images of sculptures. We always select artworks in minimal pairs of two similar images to gauge the robustness of interpretation. The complete image list can be found in Appendix A.

We present most image files directly as downloaded from Prometheus. For two paintings by Arnold Böcklin, irrelevant parts of the two images are masked out. In two other cases, recognizable writing on the images is masked to focus the model on visual information (see Appendix A for details).

The images are presented to the VLMs individually with each question. We set maximum answer length to 150 tokens and otherwise use the models' default parameters. This results in a total of 20 image pairs x 2 images x 8 questions x 3 models = 960 answers.[2]

## 4 Findings

We discuss our findings structured by observation. Figure 1 shows three artworks to illustrate the difficulties of the task[3] and Table 2 the corresponding VLM answers for the five most instructive questions. We color-code relevant passages in the VLM output as correct (green), incorrect (red), and questionable (purple).

**A Hierarchy of Skills in Image Analysis.** All three models demonstrate a good, sometimes even

---

[2] The metadata of the image files, the responses of the models, and the scripts used for the evaluation are publicly available at `https://doi.org/10.5281/zenodo.17668625`

[3] See Appendix for image source and licensing information.

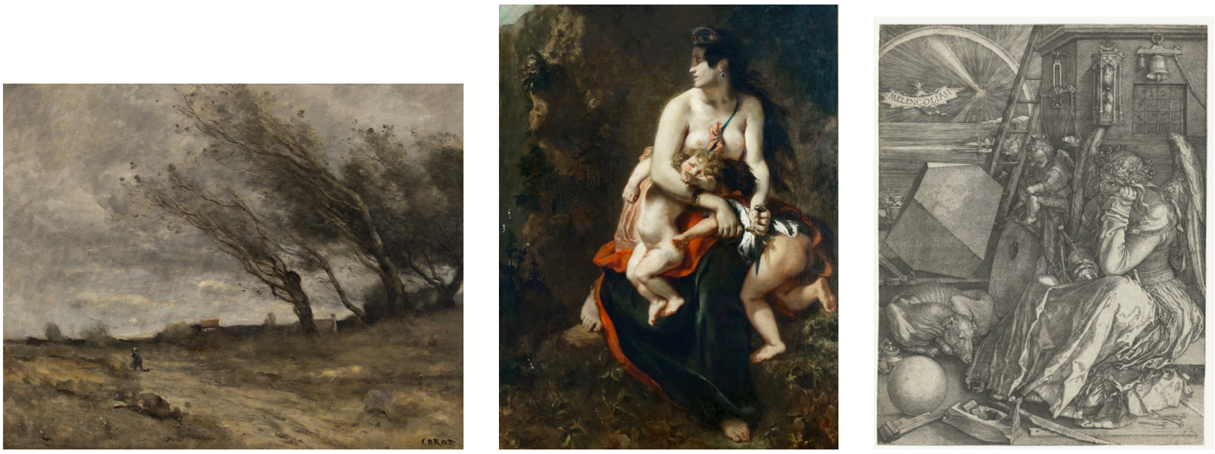


Figure 1: Examples of three difficult artworks. Left: The blast (Corot, Picture 15 in Appendix A). Center: Medea furious (Delacroix, Picture 21). Melancholia I (Dürer, Picture 27).

| Prompt | Corot (QWEN-7B) | Delacroix (QWEN-32B-AWQ) | Dürer (QWEN-7B) | Dürer (LLAVA-LLAMA-8B) |
|---|---|---|---|---|
| 2. What does this artwork show? | This artwork depicts a landscape scene with a strong wind bending the trees, creating a dynamic and dramatic effect. A solitary figure is seen walking along a path in the foreground, adding a sense of scale and human presence to the vast, open environment. The sky is filled with clouds, suggesting an overcast or stormy day. The painting captures the raw power of nature and the resilience of the human figure within it. | This artwork depicts a classical scene featuring a nude woman seated on a rocky surface, holding a child in her arms. Another child is playfully interacting with her, and the background includes lush foliage and a dark, shadowy setting, suggesting a mythological or allegorical theme. The composition and style are reminiscent of neoclassical or romantic art. | This artwork is a detailed engraving that appears to depict a scene with symbolic elements. It includes a figure in a state of repose, possibly a reclining person or a figure of death, surrounded by various objects such as a globe, a saw, and other tools. There are also two figures, one holding a baby, which could symbolize life or innocence, and another figure with wings, possibly an angel or a representation of time or fate. The presence of the clock and the hourglass suggests themes of time and mortality. [. . . ] | The artwork depicts a scene with a woman sitting on the ground, surrounded by various objects and figures that appear to be in a state of disarray or chaos. There is a ladder, a bell, a clock, and other items scattered around, and the woman seems to be in a contemplative or distressed state. The style is reminiscent of the Baroque period, known for its dramatic and intense compositions. |
| 3. Does this artwork involve an emotion? | Yes. The artwork evokes a sense of tranquility and introspection. The solitary figure walking through the landscape, combined with the gentle breeze and the expansive sky, suggests a peaceful moment of reflection. The muted colors and the soft brushstrokes contribute to a calm and contemplative atmosphere, which can be interpreted as an emotional response from the viewer. | Yes. This artwork involves emotion, as it depicts a tender and intimate moment between a woman and a child. The woman's posture and the way she holds the child suggest a sense of care, love, and protection. The overall composition, with its soft lighting and the natural setting, enhances the emotional depth of the scene, evoking feelings of maternal love and serenity. | Yes. The artwork depicts a scene with symbolic elements that can evoke emotions. The presence of the angel holding a baby, the reclining figure, and the various objects around them suggest themes of mortality, divinity, and perhaps a sense of loss or contemplation. The overall composition and the use of light and shadow can also contribute to an emotional atmosphere, inviting viewers to reflect on the deeper meanings behind the imagery. | Yes, the artwork appears to involve an emotion. The central figure, which seems to be a woman, is depicted with a contemplative or possibly sorrowful expression. Her posture and the way she holds her head suggests a moment of introspection or sadness. The overall atmosphere of the scene, with the cluttered and somewhat chaotic environment, may also contribute to a sense of emotional distress or unease. |
| 4. Is the emotion shown by this artwork a positive or a negative one? | Negative The painting depicts a stormy landscape with bent trees, suggesting strong winds and possibly rain, which can evoke feelings of unease, fear, or sadness in viewers. The solitary figure in the distance might also contribute to a sense of isolation or vulnerability [. . . ] | **Positive** The artwork depicts a serene and tender moment between a woman and a child, with the woman holding the child close and the child appearing content and secure. The overall mood is one of love, care, and tranquility, which are positive emotions. [. . . ] | Negative The artwork depicts a scene of death, with a figure lying on the ground, possibly deceased, and a dog lying beside them. The presence of an angel holding a baby suggests a theme of loss or mourning, which typically conveys a negative emotional tone. | Negative |
| 6. How is the emotion shown by this artwork expressed artistically? | The emotion in this artwork is expressed through the dynamic and dramatic posture of the trees, which appear to be bent and twisted by the wind, suggesting a sense of struggle and resilience. The muted color palette and the overcast sky contribute to a somber and contemplative mood, evoking feelings of solitude and introspection. The lone figure in the distance adds to the overall feeling of isolation and perhaps a yearning for connection or escape from the harshness of nature. The artist's use of light and shadow further enhances the emotional depth [. . . ] | The artwork expresses emotion through a combination of compositional elements, color, and body language. The central figure, a woman, is depicted with a serene and contemplative expression, her gaze directed outward, suggesting introspection or a sense of calm. The use of soft, warm lighting highlights her form and the delicate features of the child she holds, emphasizing tenderness and maternal love. The dark, shadowy background contrasts with the illuminated figures, drawing attention to their interaction and creating a sense of intimacy. [. . . ] | The emotion in this artwork is expressed through intricate details and symbolic elements. The angelic figures, one holding a baby, convey a sense of tenderness and protection. The reclining figure draped in a cloak suggests a state of rest or perhaps contemplation, while the presence of tools and a globe hints at themes of craftsmanship and exploration. The overall composition, with its detailed textures and symbolic objects like the hourglass and bell, evokes a contemplative and somewhat somber mood, reflecting on themes of time, mortality, and the passage of life. | The emotion in this artwork is expressed through the use of a somber color palette, the detailed rendering of the figures, and the overall composition. The woman, dressed in a flowing gown, sits with her head in her hands, suggesting a state of sorrow or contemplation. The surrounding environment, with its cluttered and chaotic elements, adds to the sense of distress. The use of light and shadow, as well as the positioning of the figures, contributes to the overall mood of the piece. The artist has skillfully used these elements to convey a deep emotional impact. |
| 7. Does this artwork use a symbol to visualize emotion, and if it does, how? | Yes, the artwork uses the symbol of bent trees to visualize emotion. The trees, leaning dramatically towards the left side of the canvas, suggest a strong wind or storm, evoking a sense of struggle and resilience. The figure in the foreground, seemingly small and distant, adds to the feeling of isolation and perhaps a contemplative mood amidst the powerful natural forces depicted. | Yes, this artwork uses symbols to visualize emotion. The central figure, a woman, is depicted in a contemplative or melancholic pose, holding a child who appears to be in distress or fear. The child's expression and body language, such as clutching the woman's arm, symbolize vulnerability and dependence. The dark, shadowy background and the woman's somber expression further enhance the emotional tone, suggesting themes of sorrow, protection, or maternal concern. [. . . ] | Yes, the artwork uses symbols to visualize emotion. The angel holding the baby conveys a sense of innocence and purity, while the angel with the laurel wreath and sword suggests a theme of victory or triumph over adversity, possibly symbolizing the triumph of good over evil. The presence of the dog lying down might represent loyalty or rest, adding another layer of emotional depth to the scene. The overall composition, with its detailed elements and symbolic figures, evokes a narrative that is rich in emotional and philosophical undertones. | Yes, the artwork uses a symbol to visualize emotion. The woman in the image is depicted with a halo, which is a common symbol in art to represent holiness, purity, or sainthood. This suggests that the woman is being portrayed as a virtuous or sacred figure, which can evoke feelings of reverence or admiration in the viewer. The halo's presence in the artwork serves to convey a sense of spiritual significance or moral excellence associated with the character. |

Table 2: VLM generated descriptions for the difficult artworks. Green: correct, red: incorrect, purple: questionable.

very good, ability to understand the images. However, there are significant differences in answer quality among the different question types, which roughly correlate with the order of our questions: The initial questions are answered better than those at the end. Q 1 (type of artwork) is almost always answered correctly (LLAVA-LLAMA-8B categorizes Hoepffner's black-and-white photograph as a drawing). The content descriptions (Q 2) are also mostly accurate: For Corot's 'Blast' (Fig. 1 left and Tab. 2), the model correctly identifies the motifs of the force of nature and the solitary person as well as the effects they create.

Q 3 to 5, which concern the emotional content, are still mostly answered correctly, but show a greater tendency toward evasive or inconsistent answers. As an example, consider the Corot painting, where Qwen initially speaks of a 'calm, contemplative atmosphere' and a 'peaceful moment of reflection', but answers 'negative' to the specific question about the polarity of the emotion and, with reference to the forces of nature, also brings unease, fear and vulnerability into play.

Q 6 concerning the artistic expression of emotions is answered largely well. The models identify, mostly reliably, various levels of expression of emotions, both in terms of content (composition, depiction of specific objects, facial expressions and postures of figures) and stylistic devices (color palette, brushwork). Overall, facial expressions and postures are more precisely identified than abstract forms which become the more accurate the more elements point in the same direction. For example, in Camille Corot's 'Blast' (Figure 1 left), QWEN-7B mentions the wind-blown trees and the threatening sky against which the small human figure is seen, in combination with the gloomy colors and the play of light and shadow. Dynamic compositions tend to be recognized better than static ones.

The answers to Q 7 concerning symbols are the least convincing. Here, all models frequently mention aspects of the image that do not represent symbols. Answers to questions 6 and 7 are often identical, as in the landscape paintings by Caspar David Friedrich (Pictures 9/10, cf. Appendix A), where, when asked about means of expression and symbols, QWEN-7B mentions soft and nuanced colors that evoke calm, peace, and a contemplative mood. In some cases, symbolic objects that do not appear in the paintings are also hallucinated. For example, LLAVA-LLAMA-8B calls the personification of Melancholy in Albrecht Dürer's engraving an 'angel' (Figure 1 right) because of her wings and hallucinates a halo that she does not possess. This might be due to the stylistic proximity to religious images, which often feature halos.

Comparing the answers within the minimal pairs of similar artworks, we also see a reasonable pattern: The answers are similar, making it clear that the pictures are closely related, but sufficiently dissimilar to pick out each picture's individuality.

In sum, we observe a continuum: Properties of images that can characterized concretely at the visual level are captured more reliably by the VLMs than properties requiring more interpretation. However, even these are often inferred correctly at a basic level. In our view, this ability indicates that the models were exposed to (at least some) interpretive texts on art or art history during training.

**Overextension of the concepts emotion and symbol.** Like LLMs, the VLMs consistently exhibit confirmation (or position) bias, i.e., the tendency to answer 'yes' to yes/no questions (Echterhoff et al., 2024). This applies in particular to Qs 3 and 7 (presence of emotions and emotion symbols): The models answer almost always affirmatively, even if this is wrong. Indeed, LLAVA-LLAMA-8B tends to avoid, in cases of doubt, a yes/no answer and simply lists potential symbols. The only model that gives explicit negative answers is QWEN-32B-AWQ ('This artwork does not use a symbol to visualize emotion').

As the models continue such wrongly affirmative answers, they are forced to interpret the terms 'emotion' and 'symbol' very broadly. Regarding emotions, the models refer to a small portion of the artworks in terms of emotions in the narrower sense of basic emotions (Ekman, 1999) such as *grief, fear*, or *joy*. Many others are described by general affective states – or maybe aesthetic emotions (Israeli, 1928) – such as *melancholy, loneliness*, or *awe*. For others, the models describe the emotional content with very abstract terms such as *contemplation, tranquility* and *introspection*, with only a tenuous connection to concrete visual properties. *Contemplation* in particular is used so frequently – by all models for more than half of all images – that an interpretation by reference to a (e.g., monastic) *vita contemplativa* (a philosophical life style) is unconvincing. Given that VLMs are presumably trained mostly on photographs, it appears plausible that *contemplation* characterizes the specific aesthetic quality of the works of art, in contrast to the pictures of contemporary reality that the models are more familiar with.

We see a similar situation with symbols: when clear symbols are present in the images, they are often recognized, for example, the book as a symbol of wisdom, the threatening sky and the thunderstorm as symbols of conflict and threat, or the skull as a symbol of mortality. LLAVA-LLAMA-8B does a better job in this regard; see the discussion below.

Particularly in less symbolic images, the models also tend to describe any emotion-relevant properties of the images as symbols, such as certain facial expressions or postures, the color palette, or the contrast of light and shadow. In our opinion, these are not genuine (emotion) symbols, since the emotional content here is part of the primary meaning and not added through a secondary conventionalized relationship. The models use the verb 'symbolize' very loosely to simply mean 'express'. In other places, the models identify symbols that are not conventionalized and therefore somewhat dubious. These include the interpretation of a figure standing in water as a symbol of loneliness (Whistler, Picture 33/34), or the interpretation of a cypress tree as a symbol of timelessness (Böcklin, Picture 25/26). Finally, some clear symbols, such as the dagger in Medea's hand (Figure 1 center), are overlooked when they do not fit a simple interpretation (see below).

**Lack of Consistency.** The VLMs we consider here do not possess a thought process independent of output generation. In consequence, they are unable to reflect on the consistency of their output (Marjanović et al., 2025). We observe inconsistency both within answers and across answers. For example, in justifying their answer to Q 3 (Is the artwork emotional?), the models often already identify an emotion and the form of its expression, which overlaps in content with Q 5 (specific emotion) and Q 6 (emotion expression). In addition, there are semantic relations between questions: In an ideal model, the answers to Q 4 (emotion polarity) and Q 5 (specific emotion) would always match.

We observe a correlation between consistency and the emotion's intensity (Q 8): the clearer and more intense the emotion depicted, the more consistently the models answer, similar to human behavior (Troiano et al., 2021). For images with subjective, weak, or ambivalent emotions, we often see contradictory answers to the various questions. Camille Corot's 'Blast' (Figure 1, left) falls into this category, as mentioned above. The models also frequently use hedging language ('or', 'perhaps even'). In principle, this might be an informative strategy, if it allowed the user to recognize model uncertainty. Currently, however, such hedge expressions are not used with sufficient reliability by the models: not all uncertain statements are marked as such, and even attributions that we consider unambiguous are repeatedly weakened. Thus, inconsistency is evident at this level as well.

**Easier and Harder Artwork Categories.** When comparing artworks from different periods, it is striking that the models produce the best results for representational artworks (showing recognizable objects), independently of the artistic technique. Examples include the *pietà* sculptures as well as paintings by Corot, Seurat, and Van Gogh. In comparison, the models struggle with three other categories of artworks: abstract images, allegories, and highly contextualized artworks.

There are three highly abstract images in our sample: two Deluge paintings by William Turner, two water paintings by James McNeill Whistler (Pictures 33/34), and two sections of a Mark Rothko mural (Pictures 35/36). These paintings primarily convey moods through colors and shapes, containing few concrete object. The models do recognize these moods, but due to semantic underdetermination, they tend to detect both positive and negative emotions, thus, as discussed above, offering an inconsistent interpretation. Formulations such as 'depending on the context' or 'depending on the viewer' are also offered.

The category of allegories is exemplified in our sample by the two allegories on melancholy by Albrecht Dürer (Figure 1 right) and Giovanni Benedetto Castiglione (Pictures 27/28). The models are clearly unfamiliar with the genre and cannot cope with the fact that (almost) all of the objects in the painting are to be understood symbolically. As shown in Table 2 (right), the models mention 'chaos' or describe the painting as 'overloaded'. Indeed, recognizing the image content also seems more difficult than in other categories, and QWEN-7B, in particular, hallucinates several related objects (including an angel holding a baby and a clock). As a result, the models' explanations of the emotions remain vague and contradictory, even though the general negative polarity is correctly recognized.

The third difficult category consists of artworks that require background knowledge for interpretation, such as Eugène Delacroix's painting of Medea (Figure 1 center) from Greek mythology who murders her children out of spite. QWEN-7B recognizes the relevant content (mother, children, sword, cave), the stylistic devices (contrasting colors), and also the tension depicted. In the overall interpretation, however, the model focuses on the main objects of mother and children and concludes that the image symbolizes 'balance and tenderness. . . the emotional depth of motherly love'. This may also be related to the fact that the model was (presumably) forbidden from verbalizing violent content due to its safety alignment; however, the problem also arises in other cases.

The difficult cases have in common that the emotional interpretation cannot be derived from the co-occurrence of a few objects associated with a conventionalized meaning: Either because there are no objects (abstract images), because there are too many (symbolic images), or because the conventional interpretation is overridden by a specific context (mythological images). This observation suggests that good recognition of objects and their configurations is still important for sound image interpretation. This seems plausible: the images the VLMs saw during training are likely only a fraction of works of art, and a much larger proportion are photographs, often of a journalistic nature. In these photographs, the meaning is primarily derived from the objects and only to a small extent from other properties. Furthermore, such popular or journalistic images usually operate with basic anthropological constellations (love, grief, relationships, violence) and less with complex symbols, as they are designed for broad understanding. From this perspective, it is even surprising that the VLMs are relatively successful in recognizing the stylistic and formal properties of the artworks. Again, this indicates the presence of art (historical) materials in the training data.

**Differences among VLMs.** The three models are comparable in terms of the general quality of their analysis, despite their differences in training and model sizes. They have different strengths and weaknesses, though. Table 2 (right) shows the output of LLAVA-LLAMA-8B and QWEN-32B-AWQ for the same picture for direct comparison.

The Qwen models appear to have seen more texts in their training that deal with visual arts specifically. As a consequence, their image descriptions read more professionally and often convey the content better. For example, in the landscape paining 'Morning on the Riesengebirge' (Caspar David Friedrich, Picture 10), the gentle gradation from warm yellow to cooler blue and the staggered mountain ranges are described in terms of a feeling of depth and expanse that invites the viewer to reflect on the grandeur of nature, while the sparse vegetation and the deserted landscape evoke feelings of loneliness and introspection: there is (almost) nothing to add to this from an art historical perspective. The model also recognizes other artistic means that do not operate as symbols, such as the half-open doors in Vilhelm Hammershøi's work (Pictures 31/32) that create tension. The Qwen models also has significantly more extensive factual knowledge: QWEN-7B directly recognizes some artists (Van Gogh), styles (Pointillism), and categories of artworks (*pietà*, Pictures 3–6). It derives emotional values directly from precisely recognized religious iconography (*pietà* = grief, empathy, maternal love). This knowledge also leads to correct intercultural interpretation of expressive values (Indian lady, 18th century, Picture 30).

The danger is that models rely too heavily on their knowledge of specific images. In fact, the larger QWEN-32B-AWQ model mistakes a numbers of artworks for different but related ones, for example claiming that George Seurat's 'Circus sideshow' (Picture 7) was his (better-known) 'A Sunday Afternoon on the Island of La Grande Jatte'. Consequently, the description of the painting is a mixture of what is seen in the actual image and what the model knows about the other one ('The figures in the painting are engaged in various activities. . . '). In that sense, the larger model's capacity for better memorization is in fact a liability. On the upside, however, the larger QWEN-32B-AWQ– as discussed above – is also the only one to break the confirmation bias and state correctly that some pictures are not symbolic.

In comparison, LLAVA-LLAMA-8B has less domain knowledge: the model only roughly recognizes styles and no artists. The model tends to have more difficulty interpreting complex images than the Qwen models, resorting more often to vague or attenuated expressions. A surprising strength of LLAVA-LLAMA-8B is its ability to recognize symbols in images and assign emotions better than Qwen. For example, LLAVA-LLAMA-8B recognizes the veil in Marta Hoepffner's 'Mourning' photograph (Picture 24); the book as a symbol of knowledge and wisdom in several images; the abyss as a symbol of threat; individual small figures as symbols of loneliness; and the cross and crown of thorns as symbols of the Passion – the latter even though the model has no concept of *pietà*. The model thus demonstrates a knowledge of affective symbols from different eras and cultures.

## 5 Discussion

Our case study investigated how well current vision-language models (VLMs) can be used to interpret the emotional content of artworks, presenting a dataset of 38 images to three current VLMs. By using a qualitative evaluation approach based on expert judgments, we are able to obtain a detailed profile of the VLMs' capabilities. Our results show that VLMs can recognize the content of artworks well, often also the emotions they depict and how they are expressed – largely independent of the type of artwork, their historical period, and their style. This indicates that VLMs implicitly incorporate a substantial amount of art historical knowledge on which they can build. The Qwen models demonstrate significantly more detailed knowledge and a better command of the domain language than LLAVA-LLAMA-8B.

However, VLMs continue to exhibit the typical limitations of LMs: they base their interpretations of deeper levels of meaning on observations of simple surface patterns (such as the presence of certain objects or stylistic devices) and their conventionalized meaning. This explains their ability to generalize to artworks, but fails precisely when the artist creates something novel by transcending conventionalized patterns or at least using them ambivalently. Similarly, when the patterns become too complex, as in symbolist paintings, the models fail, since globally coherent interpretations would require complex inferences (Shen et al., 2024). Given these observations, it is not surprising that the VLMs -– especially the Qwen models -– struggle with the reliable recognition of emotion symbols: The interpretation of such symbols typically arises from a long chain of inferences that determine the meaning of a symbol by intersecting the set of possible symbols, the set of interpretations of these symbols, and the set of meanings of the entire artwork.

The strength of VLMs our study identifies in dealing with conventionalized aspects of artwork already lends itself to possible concrete applications. For example, VLMs could be used in extending large catalogs of artworks with short descriptive text for better indexing. Similarly, models also seem to be sufficiently powerful for the large field of audio descriptions in e-publishing, converting images into text, for example to improve accessibility. In either case, there is still a role for experts to review and correct these results.

## 6 Limitations and Future Work

Our analysis is a case study and only considered a limited number of artworks (38) and of VLMs (3). The artworks were all drawn from the classical art history canon, including only one non-Western artwork (Picture 30) and only two artworks by women (Picture 23 and 24).

Further, our analysis of the VLM outputs follows methodological practice in art history rather than AI. It adopts a scalable reading approach, largely qualitative in nature, that does not scale well: it could not be extended easily to larger samples of artworks (to achieve a better representativeness), additional questions (to apply our approach to artwork properties other than emotions) or a larger number of raters (to minimize the impact of personal bias).

For these reasons, one important avenue for future research is the identification of evaluation approaches that combine the advantages of Ozaki et al.'s (2024) fully automatic quantitative analysis with our qualitative, more detailed analysis. A second such avenue is the fine-tuning of VLMs for the analysis of emotions in artwork. This can be hoped to improve the general ability of VLMs to recognize emotional content. Also, it would ideally curb the models' verbosity, which complicates their use for tasks like audio description generation (as mentioned above) and the field of cultural heritage in general. As usual, the challenge remains to find suitable supervision.

## References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, and C. Lawrence Zitnick. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, Santiago, Chile.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Roland Barthes. 1977. *Image, Music, Text: Essays*. Hill and Wang / Fontana Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Ernst Cassirer. 1923. *Philosophie der symbolischen Formen*, volume 1: Die Sprache. Kindler, Weimar.

Moacir P. de Sá Pereira. 2019. Mixed methodological digital humanities. In Matthew K. Gold and Lauren F. Klein, editors, *Debates in the Digital Humanities 2019*. University of Minnesota Press, Minneapolis.

Lisa Dieckmann. 2010. Prometheus: The distributed digital image archive for research and education. In Béatrice Joyeux-Prunel, editor, *L'Art et la Mesure*. Éditions Rue d'Ulm.

Jessica M. Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Paul P. Ekman. 1999. Basic emotions. In T. Dalgleish and T. Power, editors, *The Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons, Sussex, U.K.

Ute Frevert, Monique Scheer, Anne Schmidt, Pascal Eitler, Bettina Hitzer, Nina Verheyen, Benno Gammerl, Christian Bailey, and Margrit Pernau. 2011. *Gefühlswissen. Eine lexikalische Spurensuche in der Moderne*. Campus, Frankfurt am Main.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55.

Nathan Israeli. 1928. Affective reactions to painting reproductions: A study in the psychology of esthetics. *Journal of Applied Psychology*, 12(1):125–139.

Evgeny Kim and Roman Klinger. 2019. A survey on sentiment and emotion analysis for computational literary studies. *Zeitschrift für digitale Geisteswissenschaften*, 4:1–23.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages Article 47, 1–6, New York, NY, USA. Association for Computing Machinery.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*.

Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2025. DeepSeek-R1 thoughtology: Let's think about LLM reasoning. *Preprint*, arXiv:2504.07128.

Susan J. Matt. 2011. Current emotion research in history: Or, doing history from the inside out. *Emotion Review*, 3(1):117–124.

Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

Anna Näslund and Amanda Wasielewski. 2020. Cultures of digitization: A historiographic perspective on digital art history. *Visual Resources*, 36(4):339–359.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. Towards cross-lingual explanation of artwork in large-scale vision language models. In *Findings of the Association for Computational Linguistics:*

*NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.

Jan Plamper. 2012. *Geschichte und Gefühl. Grundlagen der Emotionsgeschichte*. Siedler, München.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*.

Barbara H. Rosenwein. 2010. Problems and methods in the history of emotions. *Passions in Context*, 1:1–32.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Peter N. Stearns and Carol Z. Stearns. 1985. Emotionology. clarifying the history of emotions and emotional standards. *The American Historical Review*, 90(4):813–830.

Ed S. Tan. 2000. Emotion, art, and the humanities. In M. Lewis and J. M. Haviland-Jones, editors, *Handbook of Emotions*, 2 edition, pages 116–134. Guilford Press, New York.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. Emotion ratings: How intensity, annotation confidence and agreements are entangled. In *Proceedings of the EACL WASSA workshop*, pages 50–61.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*.

Thomas Weitin. 2017. Scalable reading. *Zeitschrift für Literaturwissenschaft und Linguistik*, 47:1–6.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

## A Image List

The following list covers artists, image titles, preprocessing steps (where pertinent) and licensing information (for the images shown in Fig. 1).

| No. | Image |
|---|---|
| 1 | Charles Le Brun: *Fear (D)*<br>[Preprocessing: Writing masked] |
| 2 | Anne-Louis Girodet: *Scene from a deluge (P)* [Preprocessing: Writing masked] |
| 3 | Unknown artist: *Pietà Röttgen (S)* |
| 4 | Unknown artist: *Pietà 1875 (S)* |
| 5 | Giovanni de Fondulis: *Madonna on throne with child (S)* |
| 6 | Michelangelo: *Pietà (S)* |
| 7 | Georges Seurat: *Circus sideshow (P)* |
| 8 | Georges Seurat: *The circus (P)* |
| 9 | Caspar David Friedrich: *The Monk by the Sea (P)* |
| 10 | Caspar David Friedrich: *Morning on the Riesengebirge (P)* |
| 11 | William Turner: *Shade and Darkness – the evening of the deluge (P)* |
| 12 | William Turner: *Light and colour (Goethe's theory) – the morning after the deluge – Moses writing the book of genesis (P)* |
| 13 | Nicolas Poussin: *Landscape during thunderstorm with Pyramus and Thisbe (P)* |
| 14 | Nicolas Poussin: *Landscape with Saint Jerome (P)* |
| 15 | Camille Corot: *The blast (P)*<br>Shown in Fig. 1 (left). License: CC BY<br>Source: `https://musees-reims.fr` |
| 16 | Camille Corot: *Recollection of Mortefontaine (P)* |
| 17 | Claude Lorrain: *Harbour scene at sunset (P)* |
| 18 | Claude Lorrain: *Harbour scene with rising sun (P)* |
| 19 | Caspar David Friedrich: *View of Arkona with rising moon (D)* |
| 20 | Caspar David Friedrich: *View of Arkona by moonlight (D)* |
| 21 | Eugène Delacroix: *Medea furious (P)*<br>Shown in Fig. 1 (center). License: PDM<br>Source: Wikipedia Commons |
| 22 | William Wetmore Story: *Medea (S)* |
| 23 | Clara von Rappard: *In Trauer (P)* |

| | |
|---|---|
| 24 | Marta Hoepffner: *Ausdrucksstudie Trauer (F)* |
| 25 | Arnold Böcklin: *Villa by the sea (P)* |
| 26 | Arnold Böcklin: *Rest on the flight into Egypt (P)* [Preprocessing: Left part of picture (figures) masked] |
| 27 | Albrecht Dürer: *Melancholia I (D)*<br>[Preprocessing: Writing masked]<br>Shown in Fig. 1 (right). License: CC0<br>Source: Wikipedia Commons |
| 28 | Giovanni Benedetto Castiglione: *Melancholia (D)* |
| 29 | Henri Matisse: *Odalisque (P)* |
| 30 | Unknown artist: *Melancholy Courtesan (P)* |
| 31 | Vilhelm Hammershøi: *Interior. Study in sunlight (P)* |
| 32 | Vilhelm Hammershøi: *White doors / Open doors (P)* |
| 33 | James McNeill Whistler: *Nocturne: Blue and silver – Cremone lights (P)* |
| 34 | James McNeill Whistler: *Nocturne: Blue and silver – Chelsea (P)* |
| 35 | Mark Rothko: *Mural, Section 5 (P)* |
| 36 | Mark Rothko: *Untitled [Black on maroon (Seagram Mural)] (P)* |
| 37 | Vincent van Gogh: *Van Gogh's chair (P)* |
| 38 | Vincent van Gogh: *Gauguin's chair (P)* |

Remarks:

- The three images from Figure 1 are numbers 15, 21, and 27. Details see there.
- Artwork types are abbreviated as follows: P = painting, S = sculpture, F = photography, D = drawing
- Complete information on pictures is available at `https://doi.org/10.5281/zenodo.17668625`

# Toward Automatic Safe Driving Instruction: A Large-Scale Vision Language Model Approach

**Haruki Sakajo**[1], **Hiroshi Takato**[2] **Hiroshi Tsutsui**[2] **Komei Soda**[2,3]
**Hidetaka Kamigaito**[1], **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology (NAIST),
[2]Teatis inc., [3]Queensland university of technology
`sakajo.haruki.sd9@naist.ac.jp`
`{tak, james}@dotsfty.com` `komei.soda@connect.qut.edu.au`
`{kamigaito.h, taro}@is.naist.jp`

## Abstract

Large-scale Vision Language Models (LVLMs) exhibit advanced capabilities in tasks that require visual information, including object detection. These capabilities have promising applications in various industrial domains, such as autonomous driving. For example, LVLMs can generate safety-oriented descriptions of videos captured by road-facing cameras. However, ensuring comprehensive safety requires additional monitoring driver-facing views to detect risky events, such as the use of mobiles while driving. Thus, the ability to process synchronized inputs is necessary from both driver-facing and road-facing cameras. In this study, we develop a model integrating two video inputs and investigate the capabilities of LVLMs by constructing a dataset and evaluating their performance on this dataset. Our experimental results demonstrate that while pre-trained LVLMs have limited effectiveness, fine-tuned LVLMs can generate accurate and safety-aware driving instructions. Nonetheless, several challenges remain, particularly in detecting subtle or complex events in the video. Our findings and error analysis provide valuable insights that can contribute to the improvement of LVLM-based systems in this domain.

## 1 Introduction

The promising capabilities of Large Language Models (LLMs) are changing this society by assisting various tasks, e.g., coding (Rozière et al., 2024) and education (Liu et al., 2024). Large-scale Vision Language Models (LVLMs) possess high capabilities in the intersection of vision and language tasks, leveraging the capabilities of LLMs, such as inference and instruction following, by integrating a vision encoder. Therefore, LVLMs have been adopted across domains that require both visual and textual information, including the medical application (Li et al., 2023a; Yan et al., 2024; Pal and Sankarasubbu, 2024) and driving assistance (Arai et al., 2025; Duan et al., 2024; Xuan et al., 2024).

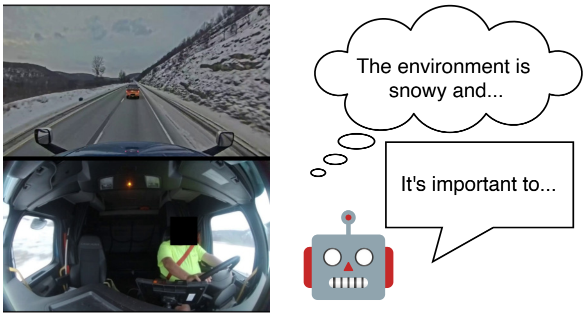

[redacted]

Figure 1: Illustration of an application of this study. A Model provides driving instructions for the given video.

In a driving domain, LVLMs are used to detect objects on the road, generate actions, and provide safe driving instructions (Zhou et al., 2024; Lu et al., 2025; Duan et al., 2024; Xuan et al., 2024). With the rapid growth of the dashcam industry, LVLMs are increasingly exploited to interpret driving scenes captured by the dashcam. Previous studies investigated the capabilities of LVLMs to interpret vehicle behavior and suggest actions for safe driving (Duan et al., 2024; Xuan et al., 2024). However, while a single dashcam for the road-facing view is enough to detect risky actions of vehicles (e.g., harsh turns), a driver-facing view using an additional dashcam is needed, for example, when detecting the driver's use of mobiles while driving.

In this study, we develop LVLM-based models that generate instructions for safe driving based on two video streams that synchronize driver-facing and road-facing views. We collect such videos and annotate each clip in two stages: first, by detecting events that happened in them, and second, by generating safe driving instructions. We create a conversation-style dataset, where a user asks what is happening in the video, followed by a request to generate safe driving instructions. Our experiments demonstrate that fine-tuned LVLMs on our dataset can generate safe driving instructions that are well-aligned with the visual context, whereas pre-trained

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 13–24
December 23, 2025 ©2025 Association for Computational Linguistics

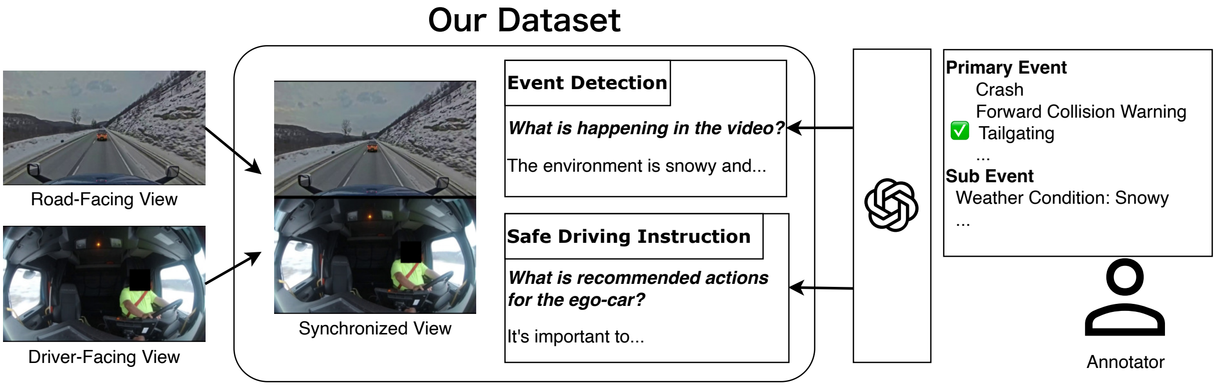


Figure 2: Overview of our dataset construction approach. The dataset contains videos with synchronized driver-facing and road-facing views. GPT-4o generates the gold answers to event detection and safe driving instruction questions based on annotated labels for each video.

| | |
|---|---|
| Human | What is happening in the video? |
| Assistant | The environment is icy with clear weather as you approach an intersection. A rolling stop occurs as you navigate the intersection at 8 mph while properly using your turn signal. |
| Human | What is recommended actions for the ego-car? |
| Assistant | \ud83d\udca1Paying extra attention to road conditions due to ice is important. Make sure to bring your vehicle to a complete stop at intersections to ensure safety, especially in icy conditions where stopping distances are longer. |

Table 1: An example of the question and its answer. The top is the event detection, and the bottom is the safe driving instruction.

models generate generic guidance. Nonetheless, even the fine-tuned models occasionally fail to detect risky events (e.g., harsh turns). Our findings and analysis provide valuable insights that help improve LVLM-based driving instruction systems, as illustrated in Figure 1.

## 2 Background and Related Work

### 2.1 Large-scale Vision Language Models

Large-scale Vision Language Models (LVLMs), such as Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b), and LLaVA (Liu et al., 2023), integrate a vision encoder with a Large Language Model (LLM), enabling them to process visual inputs (e.g., images) and achieving advanced performance on Visual Question–Answering (VQA) tasks (Liu et al., 2023; Bai et al., 2025). Visual instruction tuning is also effective for further performance improvement (Liu et al., 2023). The enhanced capabilities of LVLMs are helpful across various domains, including disease detection from clinical images (Yan et al., 2024; Pal and Sankarasubbu, 2024), artwork explanation (Hayashi et al., 2024; Ozaki et al., 2025), and vowel prediction from MRI-based articulatory images and videos (Sakajo et al., 2025).

### 2.2 Language Models in Driving Scenario

LLMs and LVLMs are also helpful in driving domains, and several studies have demonstrated their capabilities (Zhou et al., 2024). For example, LVLMs visually understand traffic signs (Lu et al., 2025), and LLMs can be utilized to develop models for processing LiDAR data (Yang et al., 2023) and autonomous driving (Xu et al., 2024). LVLMs are also leveraged for constructing a driving dataset (Arai et al., 2025). In the AI City Challenge (Wang et al., 2024), LVLM-based approaches (Duan et al., 2024; Xuan et al., 2024) achieved advanced performance for the traffic safety description and analysis task, suggesting that LVLMs have the capabilities to learn and provide descriptions regarding safe driving. However, the capabilities of LLMs to instruct safe driving with synchronized driver-facing and road-facing RGB cameras are unexplored.

## 3 Dataset Construction

We constructed a dataset to evaluate LVLMs in the context of safe driving instructions, as illustrated in Figure 2. The dataset comprises videos and questions, along with expert-annotated answers.

| **Primary Event** | **Description** |
|---|---|
| Crash | Any visible crash involving ego-vehicle or other vehicles. |
| Forward Collision Warning | An event where the ego-vehicle is at risk of imminent front-end collision. |
| Tailgating (Following Distance) | Following another vehicle at a dangerously short distance. |
| Harsh Brake | A sudden, strong deceleration by the ego-vehicle. |
| Harsh Turn | A sharp, abrupt turn indicating potential loss of control or evasive maneuver. |
| Rolling Stop | Failure to come to a complete stop at stop signs or similar control points. |
| Mobile Usage | The driver is observed using a mobile phone (hands-on or hands-free). |
| Inattentive Driving | Observable distraction or loss of attention by the driver. |
| Lane Departure | Vehicle crosses out of its lane without clear intention or necessity. |
| Other Events | Includes seatbelt violations or distraction events. |

Table 2: Primary Event Categories for Annotation.

| **Sub-Event** | **Options** |
|---|---|
| Lane Cut Off | Proper use of turn signal,Improper use of turn signal |
| Lane Change | Proper use of turn signal, Improper use of turn signal, To avoid primary event, Root cause of primary event |
| Turn (Other Vehicles) | Proper use of turn signal, Improper use of turn signal |
| Turn (Ego Vehicle) | Proper use of turn signal (listen to the audio), Improper use of turn signal |
| Signs of Aggressive Reaction | Vehicle maneuver, Aggressive language, Honk, None, Unknown |
| Signs of Distraction | Smoking, Mobile phone, Playing with hair, Drinking, Eating, Picking something from the floor, Reaching behind the backseat, Yawning, None, Unknown |
| Weather Condition | Clear, Rainy, Foggy, Snowy |
| Road Condition | Dry, Wet, Icy |
| Visibility Condition | Clear, Poor |
| Road Information | Highway, Highway merge, Local road, Intersection, 3-leg intersection, School zone, Construction zone, Residential area, Rural roads, Tunnel, Pedestrian crossroad |
| Speed Management | Decrease, Maintain, Increase |

Table 3: Sub-Event Categories and Options.

| | Train | Validation | Test |
|---|---|---|---|
| Samples | 1,719 | 215 | 215 |
| Duration (s) | 18,720 | 2,311 | 2,371 |
| Frames | 561,223 | 69,291 | 72,836 |

Table 4: Dataset statistics.

### 3.1 Task

We evaluate LVLMs' capabilities and challenges using a conversation-style VQA task related to safe driving. We provide LVLMs with synchronized driver-facing and road-facing videos captured using RGB cameras and then ask the LVLMs to explain what happens in the video and generate instructions for safe driving.

### 3.2 Video Collection

We collected vehicle speed and video recordings from both driver-facing and road-facing RGB cameras and lined them up vertically as unified clips, placing the road-facing view on the top and the driver-facing view on the bottom. Each video in our dataset presents synchronized views of both the driver and the road.

| Parameter | Value |
|---|---|
| Batch size | 8 |
| Epoch | 3 |
| Learning rate | 1e-5 |
| Learning rate scheduler | cosine |
| Adam $\beta_1$ | 0.9 |
| Adam $\beta_2$ | 0.999 |
| Adam $\epsilon$ | 1e-8 |
| Precision | BF16 |
| Video Max Pixels | 16,384 |
| Video Min Pixels | 256 |
| Video Maxlen | 128 |
| Video FPS | 2 |
| Seed | 42 |

Table 5: Hyperparameters for fine-tuning.

### 3.3 Question Definition

We adopt a Chain-of-Thought (Kojima et al., 2022) and a conversation-style template to facilitate step-by-step reasoning. The dataset has two questions: (1) *"What is happening in the video?"* and (2) "*What is recommended actions for the ego-car?"*. We refer to the first type as **event detection** questions/answers and the second as **safe driving instruction** questions/answers. An example of each type of question–answer pair is presented in Table 1.

This Chain-of-Thought format is designed to guide the model through a reasoning process that first identifies events in the video and then infers appropriate driving actions based on those observations.

### 3.4 Annotations

To support structured, context-rich labeling of driving scenarios, we implemented a three-step process.

**Primary event selection.** An annotator begins by selecting a single **primary event** from a predefined taxonomy of safety-critical driving behaviors, as listed in Table 2. These events capture the core nature of the incident.

**Sub-event selection.** Next, an annotator is encouraged to select as many relevant **sub-events** as necessary to describe the contributing context. These sub-events, summarized in Table 3, include surrounding vehicle behaviors (e.g., lane changes, turn maneuvers), environmental conditions (e.g., weather, visibility), or behavioral cues (e.g., signs of aggression). This multi-label scheme enables fine-grained characterization of complex traffic scenes.

**Summary generation.** In the final step, a natural language annotation is automatically generated using GPT-4o (OpenAI et al., 2024). The model takes as input the selected primary and sub-events, along with auxiliary data such as the vehicle's speed at the time of the event. Based on this information, GPT-4o generates a descriptive summary that answers two key questions: "What is happening in the video?" and "What is recommended actions for the ego-car?" Finally, experts manually review the generated descriptions and confirm the quality.

### 3.5 Data Statistics

Table 4 shows our dataset statistics. A video has an approximate duration of 10 seconds and 30 frames per second. The number of primary events and sub-event options is provided in Appendix A.

## 4 Experimental Settings

### 4.1 Dataset

We use our dataset introduced in Section 3 and treat the videos as two frames per second.

### 4.2 Models

In this study, we utilize Qwen2.5-VL (Bai et al., 2025) 3B and 7B models, which are available even with limited computational resources. We also fine-tune these models on our dataset. We refer to fine-tuned Qwen2.5-VL-3B and fine-tuned Qwen2.5-VL-7B as Qwen2.5-VL-3B (FT) and Qwen2.5-VL-7B (FT), respectively.

### 4.3 Training and Inference

Models receive instruction and video inputs, while auxiliary sensor data, e.g., vehicle speed, was incorporated during dataset construction. This approach reflects the practical consideration that dashcams are easily deployable, whereas sensor installation requires additional costs.

**Training.** We freeze the vision encoder and train only the language model with full-parameter supervised fine-tuning. LVLMs fine-tuning is conducted using the LLaMA-Factory (Zheng et al., 2024) with DeepSpeed ZeRO stage 2 (Rajbhandari et al., 2020). We fine-tuned LVLMs using eight NVIDIA A100-SXM4-40GB GPUs and used LLaMA-Factory version 0.9.2.dev0 with minor modifications to load models correctly. Table 5 provides the hyperparameters.

**Inference.** We test LVLMs and fine-tuned LVLMs under the zero-shot setting. The evaluations are performed on an NVIDIA L4 GPU.

### 4.4 Metrics

We evaluate the quality of generated text by comparing it to the reference text in the dataset using BERTScore (Zhang* et al., 2020) and BLEU socres (Papineni et al., 2002) as evaluation metrics. We use the original implementation[1] for BERTScore using RoBERTa (Liu et al., 2019) and sacreBLEU (Post, 2018)[2] for BLEU scores.

## 5 Result and Discussion

Table 6 shows the results of each model on our dataset, and Tables 7 and 8 show the samples of generated text for event detection and safe driving instruction. Before fine-tuning, Qwen2.5-VL-3B performs better in terms of F1 score on BERTScore for the safe driving instruction than Qwen2.5-VL-7B, while the 7B model outperforms the 3B model in event detection. This suggests that the parameter size is irrelevant to the task performance of pre-trained models. Fine-tuning improves both BERT scores and BLEU scores, indicating that models can learn this task correctly.

---

[1] `https://github.com/Tiiiger/bert_score`

[2] `https://github.com/mjpost/sacrebleu`

| Model | Event Detection |  |  |  | Safe Driving Instruction |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | P | R | F1 | BLEU | P | R | F1 | BLEU |
| Qwen2.5-VL-3B*† | 0.8180*† | 0.8504*† | 0.8337*† | 1.44*† | 0.8273*† | 0.8590*† | 0.8426*† | 1.66*† |
| Qwen2.5-VL-7B | 0.8235*† | 0.8498*† | 0.8364*† | 1.26*† | 0.8086*† | 0.8662*† | 0.8364*† | 1.15*† |
| Qwen2.5-VL-3B (FT) | 0.8901*† | 0.9041*† | 0.8970*† | 10.59* | 0.8935*† | 0.8954*† | 0.8944*† | 11.83*† |
| Qwen2.5-VL-7B (FT) | 0.8999*† | 0.9070*† | 0.8983*† | 10.24* | 0.8994*† | 0.9002*† | 0.8997*† | 13.40*† |

Table 6: Results of event detection and safe driving instruction tasks. Metrics include BERTScore Precision (P), Recall (R), F1, and BLEU. We performed statistical significance tests using 10,000 bootstrap resamples with a significance level of $p = 0.05$. * and † indicate statistically significant differences between fine-tuned and non-fine-tuned conditions and across models, respectively. The performance differences across models and between fine-tuning and non-fine-tuning conditions are statistically significant, except for the BLEU scores for the event detection between Qwen2.5-VL-3B (FT) and Qwen2.5-VL-7B (FT)

| Question | What is happening in the video? |
|---|---|
| Ground-truth | The environment is icy with clear weather as you approach an intersection. A rolling stop occurs as you navigate the intersection at 8 mph while properly using your turn signal. |
| Qwen2.5-VL 3B | The video shows a split-screen view of a snowy street scene and the interior of a vehicle. The top half of the screen captures the outside view, showing a snow-covered road with tire tracks, buildings on either side, and a stop sign ahead. The bottom half of the screen shows the inside of the vehicle, where a person is driving. The driver appears to be looking straight ahead, and there is a piece of paper or document on the dashboard. The scene suggests that the driver is navigating through a snowy environment, possibly preparing to make a turn or stop at the stop sign.<br>[BERTScore] P: 0.8202, R: 0.8554, F: 0.8374 [BLEU] 0.5540 |
| Qwen2.5-VL 7B | The video shows a first-person perspective from inside a vehicle, likely a truck or bus, as it drives through a snowy town. The view alternates between the road ahead and the interior of the vehicle, where a person is driving. The road appears to be partially cleared of snow, with patches of snow on the sides. There are buildings on both sides of the road, some of which have signs indicating they are businesses. The driver is focused on the road, and there are other vehicles visible in the background, including a white car and an orange vehicle parked near a building labeled “Body Shop.” The weather seems overcast, and the ground is covered with snow.<br>[BERTScore] P: 0.8139, R: 0.8439, F: 0.8286 [BLEU] 0.4693 |
| Qwen3.5-VL 3B (FT) | The weather is clear and the road conditions are dry as you approach an intersection. You perform a rolling stop at the intersection, maintaining a speed of 10 mph. There are no signs of distraction or aggressive reaction from your side.<br>[BERTScore] P: 0.8997, R: 0.9149, F: 0.9072 [BLEU] 14.3398 |
| Qwen2.5-VL 7B (FT) | The environment around you is snowy with icy road conditions, making it challenging to drive safely. You approach an intersection and perform a rolling stop at 10 mph while using aggressive language.<br>[BERTScore] P: 0.9132, R: 0.9202, F: 0.9167 [BLEU] 7.6260 |

Table 7: Sample ground-truth and generated sentences of Qwen2.5-VL 3B in event detection with evaluation metrics include BERTScore Precision (P), Recall (R), F1, and BLEU.

### 5.1 Result Comparison

**Parameter size influence.** The difference in the F1 score of BERTScore between Qwen2.5-VL-3B and Qwen2.5-VL-7B on the event detection was not significant, indicating that the performance of the event detection task is saturated in the zero-shot setting. The results demonstrate that the 3B model performs better for safe driving instruction than the 7B model, which can be attributed to the output

| Question | What is recommended actions for the ego-car? |
|---|---|
| Ground-truth | \ud83d\udca1Paying extra attention to road conditions due to ice is important. Make sure to bring your vehicle to a complete stop at intersections to ensure safety, especially in icy conditions where stopping distances are longer. |
| Qwen2.5-VL 3B | Based on the video description, the recommended action for the ego-car (the vehicle being driven) would be to follow the instructions given by the driver. Since the driver is looking straight ahead and appears to be focused on the road, it is likely that they have made a decision about the next move. The driver may be preparing to make a turn or stop at the stop sign. Therefore, the ego-car should follow the driver's instructions and navigate accordingly. This could involve accelerating, decelerating, or changing lanes as directed by the driver.<br>[BERTScore] P: 0.8356, R: 0.8400, F: 0.8378 [BLEU] 1.0798 |
| Qwen2.5-VL 7B | In the context of the video, the ego-car (the vehicle being driven) is navigating a snowy environment. Here are some recommended actions that could help ensure safe driving:<br>1. **Drive Slowly**: Snowy roads can be slippery, so reducing speed helps maintain better control of the vehicle.<br>2. **Maintain a Safe Following Distance**: Increase the distance between your vehicle and the one in front to allow more time to react to any sudden stops or changes in traffic conditions.<br>3. **Use Low Beams**: If visibility is reduced due to snow or ice, use low beams to improve visibility without causing glare for other drivers.<br>4. **Avoid\u6025\u8f6c\u5f2f (Sharp Turns)**: Snow can cause the tires to lose traction, making sharp turns particularly dangerous. Try to avoid them if possible.<br>5. **Check Tires**: Ensure that your tires have adequate tread depth and are properly inflated. Snow chains may also be necessary in very snowy conditions.<br>6. **Be Aware of Other Vehicles**: Watch out for other vehicles, especially those that might be struggling to navigate the snow. Be prepared to stop or slow down if necessary.<br>7. **Stay Alert**: Snowy conditions can be deceptive, and visibility might be reduced. Stay alert and be ready to react to any changes in the road or traffic.<br>8. **Use Turn Signals**: Signal your intentions clearly to other drivers, especially when changing lanes or turning.<br>9. **Keep a Full Tank of Fuel**: Snowy conditions can lead to unexpected delays, so having enough fuel is crucial.<br>10. **Plan Your Route**: If possible, plan your route in advance to avoid areas with heavy snow accumulation or icy patches.<br>By following these recommendations, the ego-car can navigate the snowy conditions safely and efficiently.<br>[BERTScore] P: 0.8089, R: 0.8662, F: 0.8366 [BLEU] 0.4625 |
| Qwen2.5-VL 3B (FT) | While it's good that you're paying attention to the road, it's important to come to a complete stop at intersections to ensure safety and compliance with traffic laws. Practice making full stops to improve your driving skills and enhance safety.<br>[BERTScore] P: 0.9062, R: 0.8896, F: 0.8978 [BLEU] 22.1907 |
| Qwen2.5-VL 7B (FT) | \ud83d\udca1In icy and snowy conditions, it's crucial to come to a complete stop at intersections to ensure safety. Try to remain calm and avoid aggressive language as it can distract from the primary task of driving safely.<br>[BERTScore] P: 0.9142, R: 0.9243, F: 0.9192 [BLEU] 22.2504 |

Table 8: Sample ground-truth and generated sentences of Qwen2.5-VL 3 in safe drivinginstruction with evaluation metrics include BERTScore Precision (P), Recall (R), F1, and BLEU.

tendencies in the 7B model. The 7B model outputs general recommendations for safe driving instruction before fine-tuning, as shown in Table 9, resulting in lower precision and higher recall. For further analysis, we also computed self-BLEU scores (Zhu et al., 2018) for each event using the outputs generated by each model to assess diversity. The self-BLEU scores, as shown in Table 10, also indicate

| Model & Event Type | Top-10 4-gram words |
|---|---|
| Qwen2.5-VL-3B | |
| Event Detection | half of the video; The video shows a; top half of the; The video shows two; the interior of the; The top half of; bottom half of the; The bottom half of; The interior of the; interior of the vehicle. |
| Safe Driving Instuction | Based on the video; on the video description,; the ego-car (the vehicle; for the ego-car (the; there are no specific; are no specific actions; recommended actions for the; the video description, there; video description, there are; a safe distance from. |
| Qwen2.5-VL-3B (FT) | |
| Event Detection | no signs of distraction; signs of distraction or; of distraction or aggressive; There are no signs; are no signs of; The footage shows you; footage shows you driving; distraction or aggressive reaction; or aggressive reaction from; aggressive reaction from your. |
| Safe Driving Instruction | While it's good that; come to a complete; to a complete stop; a complete stop at; it's good that you're; to come to a; it's important to come; important to come to; complete stop at intersections; stop at intersections to. |
| Qwen2.5-VL-7B | |
| Event Detection | The video appears to; half of the screen; of the screen shows; the screen shows the; video appears to be; the interior of the; appears to be a; shows the interior of; to be a split-screen; be a split-screen view. |
| Safe Driving Instruction | In the context of; the context of the; are some general recommendations; some general recommendations for; the ego-car (the vehicle; context of the video,; a safe distance from; ego-car (the vehicle being; distance from the vehicle; for the ego-car (the. |
| Qwen2.5-VL-7B (FT) | |
| Event Detection | signs of distraction or; of distraction or aggressive; no signs of distraction; There are no signs; are no signs of; clear weather and dry; weather and dry road; The footage shows you; footage shows you driving; and dry road conditions. |
| Safe Driving Instruction | come to a complete; to a complete stop; a complete stop at; increase your following distance; maintain a safe following; your following distance to; to maintain a safe; a safe following distance; important to maintain a; safe following distance to. |

Table 9: The top 10 4-grams in each response.

| | Detection | Instruction |
|---|---|---|
| Qwen2.5-VL-3B | 84.4359 | 82.8134 |
| Qwen2.5-VL-7B | 83.6289 | 85.7856 |
| Qwen2.5-VL-3B (FT) | 96.4625 | 95.8846 |
| Qwen2.5-VL-7B (FT) | 95.5442 | 93.5108 |

Table 10: Self-BLEU scores. "Detection" and "Instruction" denote "Event Detection" and "Safe Driving Instruction", respectively.

that the 7B model outputs less diverse texts for the safe driving instruction when compared with the 3B model.

**Performance improvement by fine-tuning.** Fine-tuning improves overall performance, and Qwen2.5-VL-7B (FT) outperforms Qwen2.5-VL-3B (FT) on both tasks in terms of BERTScore, while Qwen2.5-VL-3B outperforms Qwen2.5-VL-7B before fine-tuning. Figures 3, 4 5 and 6 also show that fine-tuning improves overall performance. On the other hand, Table 6 shows that the difference in the BLEU scores between both fine-tuned models on the event detection task is not significant. This suggests that a larger parameter size has a positive effect on the fine-tuning performance of LVLMs for this task. In contrast, the final performance after fine-tuning remains consistent across model sizes with respect to BLEU scores.

## 5.2 Error Analysis

We focus on the subset of samples for which BERTScore of the safe driving instruction falls within the bottom 25%. Approximately 4% of all the samples are shared across the bottom 25% subsets for all models, which we refer to as the "difficult subset". Within this subset, 33% of the samples are annotated as good driving, and another 33% involve scenarios where the ego-car is turning right and left. Although all models generated recommendations to improve already good driving behaviors, the suggestions for safer driving varied slightly, resulting in relatively low scores.

For the turning right and left scenarios, the gold answers typically recommend turning while reducing speed. However, even fine-tuned models produced irrelevant suggestions, such as mentioning a stop sign not presented in the video. These observations suggest that while fine-tuned models are capable of generating various safety-related suggestions, they still struggle to detect issues such as excessive speed during turns.

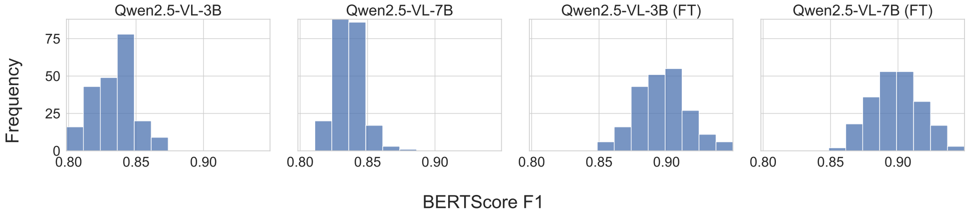


Figure 3: Score distribution of BERTScore F1 of event detection.

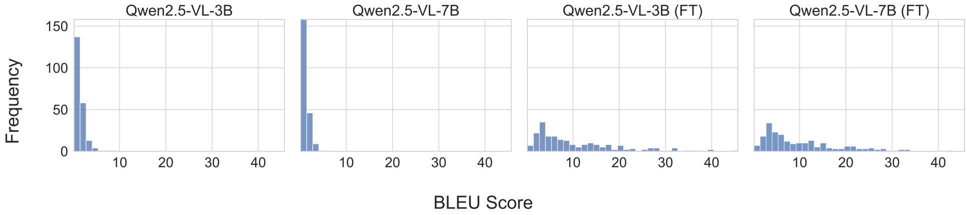


Figure 4: Score distribution of BLEU of event detection.

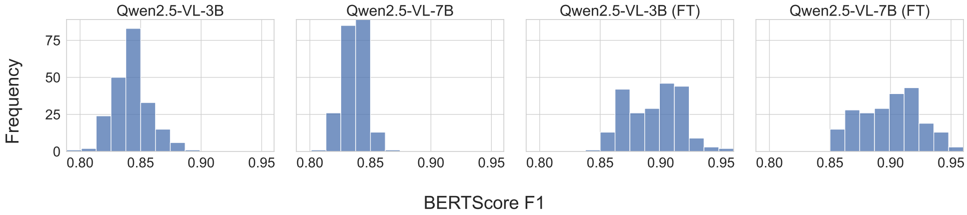


Figure 5: Score distribution of BERTScore F1 of safe driving instruction.

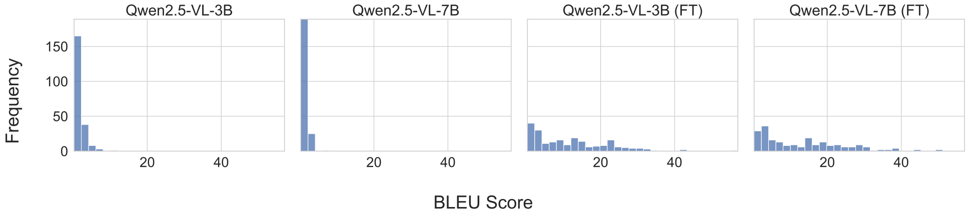


Figure 6: Score distribution of BLEU of safe driving instruction.

In the difficult subset, approximately 10% have errors related to the driver-facing view, where a driver holds and uses a phone while keeping their eyes on the road. This might suggest that LVLMs can provide safe driving instruction regarding drivers' behaviors, while they struggle to generate it regarding vehicle behaviors. This phenomenon is explained by the relative ease with which LVLMs can detect a driver holding an object, as opposed to estimating vehicle speed, which requires more temporal reasoning.

### 5.3 Unimodal Biases

As discussed in Section 5.1, in several cases, pretrained models provide general suggestions regardless of the videos, as shown in Table 8. We also observed that the fine-tuned models mentioned objects not presented in the video in Section 5.2. This behavior can be attributed to unimodal biases, specifically language biases (Goyal et al., 2017; Agrawal et al., 2018; Zhu et al., 2020; Abbasnejad et al., 2020; Chen et al., 2024), where models' outputs are biased toward textual information in the given inputs. However, while language biases have been attributed to the model's learning of the relationships between question-answer pairs in the training data (Agrawal et al., 2018), it is unclear whether the training data includes question-answer pairs regarding safe driving instruction. Given that Sakajo et al. (2025) observed the model generating identical answers for different images with the same question in a phonetics-related VQA task, our findings suggest that language priors emerge in certain domains.

### 5.4 Task Difficulty and Application Possibility

The results reveal that this task is challenging for LVLMs without fine-tuning, whereas fine-tuning improves performance. Our error analysis in Section 5.2 also indicates that several failure cases happen for good driving videos, and suggestions for safer driving vary slightly. Those discussions suggest that our fine-tuned models can be applied to safe driving instruction systems, although several challenges remain in certain situations, such as instructing against a harsh turn.

## 6 Conclusion

In this study, we constructed a dataset comprising synchronized driver-facing and road-facing video streams, along with step-by-step question–answer pairs. We fine-tuned LVLMs on our dataset for safe driving instructions and investigated their capabilities and performance in detecting risky events and providing safe driving instructions. Our experimental results reveal that fine-tuned LVLMs demonstrate the capability of suggesting safety-aware driving instructions, while detecting several events remains challenging, even for fine-tuned models. Our findings suggest that LVLMs can be safe driving instructors, although there is room for improvement.

## Limitations

**Dataset size.** As described in Section 3, our dataset comprises 1,719 training samples, 215 validation samples, and 215 test samples, which can be considered relatively small in scale. However, the collection of synchronized driver-facing and road-facing views requires a complicated setup, characterizing this task as a low-resource scenario. In this study, we investigated model performance using the current dataset as an initial step, with evaluation on a larger dataset left for future work.

**Dataset quality.** The instructions in our dataset were generated using GPT-4o, which might raise concerns regarding their quality. However, as detailed in Section 3, the generated texts were manually reviewed to ensure the quality.

**Model selection.** In this study, we selected two base models: Qwen2.5-VL-3B and Qwen2.5-VL-7B. While this choice might constrain our investigation of the scaling law in this task and performance variation across models, it remains justifiable. The Qwen2.5-VL series achieves advanced performance on various benchmarks, including Video-MME (Fu et al., 2024), and our objective is to evaluate model effectiveness for driving instruction. Accordingly, focusing on the Qwen2.5-VL series and its relatively small variants is appropriate for our investigation.

## Ethical Considerations

Our dataset contains videos that capture drivers. We collect these videos legitimately and use them within the prescribed scope.

## Acknowledgement

We thank the anonymous reviewers for their valuable comments and suggestions. This project is supported by Teatis inc. through the provision of the dataset and computational resources.

## References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10041–10051.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. 2025. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. In *Findings of the Association for Computational*

*Linguistics: EMNLP 2024*, pages 16449–16469, Miami, Florida, USA. Association for Computational Linguistics.

Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. 2024. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7180–7189.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729, Bangkok, Thailand. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: bootstrapping language-image pretraining with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J. Malan. 2024. Teaching cs50 with ai: Leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 750–756, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. 2025. Can lvlms obtain a driver's license? A benchmark towards reliable AGI for autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5838–5846.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. Towards cross-lingual explanation of artwork in large-scale vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3773–3809, Albuquerque, New Mexico. Association for Computational Linguistics.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Gemini goes to Med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 21–46, Mexico City, Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron

Grattafiori, Wenhan Xiong, Alexandre Défossez, and 7 others. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Haruki Sakajo, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Tonguescape: Exploring language models understanding of vowel articulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12605–12619, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, and 5 others. 2024. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193.

Khai Trinh Xuan, Khoi Nguyen Nguyen, Bach Hoang Ngo, Vu Dinh Xuan, Minh-Hung An, and Quang-Vinh Dinh. 2024. Divide and conquer boosting for enhanced traffic safety description and analysis with large vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7046–7055.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024. Worse than Random? An Embarrassingly Simple Probing Evaluation of Large Multimodal Models in Medical VQA. In *GenAI for Health: Potential, Trust and Policy Compliance*.

Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. 2023. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *Preprint*, arXiv:2312.14074.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. 2024. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, pages 1–20.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI-2020, page 1083–1089. International Joint Conferences on Artificial Intelligence Organization.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

## A Dataset Statistics (Detail)

Table 11 shows the number of events or options in each dataset split.

| | Train | Val. | Test |
|---|---|---|---|
| **Primary Events** | | | |
| Crash | 3 | 0 | 0 |
| Forward Collision Warning | 41 | 6 | 7 |
| Tailgating | 90 | 21 | 7 |
| Harsh Brake | 253 | 20 | 26 |
| Harsh Turn | 15 | 4 | 2 |
| Rolling Stop | 308 | 42 | 42 |
| Mobile Usage | 87 | 14 | 11 |
| Inattentive Driving | 142 | 17 | 20 |
| Lane Departure | 0 | 0 | 0 |
| **Sub Events** | | | |
| Lane Cut Off | | | |
| Improper use of turn signal | 18 | 3 | 0 |
| Proper use of turn signal | 21 | 7 | 1 |
| Lane Change | | | |
| Improper use of turn signal | 11 | 0 | 1 |
| Proper use of turn signal | 84 | 11 | 9 |
| To avoid primary event | 21 | 2 | 3 |
| Root cause of primary event | 24 | 3 | 2 |
| Turn (Other Vehicles) | | | |
| Proper use of turn signal | 10 | 1 | 3 |
| Improper use of turn signal | 3 | 0 | 0 |
| Turn (Ego Vehicle) | | | |
| Proper use of turn signal | 114 | 15 | 20 |
| Improper use of turn signal | 102 | 18 | 10 |
| Signs of Aggressive Reaction | | | |
| Vehicle maneuver | 1 | 0 | 0 |
| Aggressive language | 18 | 1 | 2 |
| Honk | 9 | 2 | 0 |
| None | 634 | 72 | 85 |
| Unknown (Dashcam Issue) | 209 | 30 | 25 |
| Signs of Distraction | | | |
| Smoking | 26 | 3 | 5 |
| Mobile phone | 18 | 3 | 2 |
| Playing with hair | 2 | 1 | 0 |
| Drinking | 14 | 0 | 2 |
| Eating | 18 | 2 | 3 |
| Picking something from floor | 2 | 0 | 1 |

| | | | |
|---|---|---|---|
| Reaching behind backseat | 0 | 1 | 0 |
| Yawning | 1 | 0 | 0 |
| None | 580 | 63 | 75 |
| Unknown | 218 | 32 | 25 |
| Weather Condition | | | |
| Clear | 1380 | 184 | 172 |
| Rainy | 92 | 6 | 10 |
| Foggy | 0 | 1 | 0 |
| Snowy | 60 | 7 | 9 |
| Road Condition | | | |
| Dry | 1362 | 181 | 170 |
| Wet | 98 | 6 | 11 |
| Icy | 72 | 11 | 11 |
| Visibility Condition | | | |
| Clear | 0 | 0 | 0 |
| Poor | 0 | 0 | 0 |
| Road Information | | | |
| Highway | 893 | 124 | 102 |
| Highway merge | 22 | 3 | 2 |
| Local Road | 8 | 0 | 0 |
| Intersection | 348 | 28 | 46 |
| 3-Leg intersection | 166 | 31 | 26 |
| School zone | 0 | 1 | 0 |
| Construction Zone | 10 | 3 | 2 |
| Residential area | 66 | 8 | 14 |
| Rural roads | 12 | 2 | 1 |
| Tunnel | 1 | 0 | 0 |
| Pedestrian crossroad | 22 | 2 | 1 |
| Parking | 27 | 2 | 6 |
| Speed Management | | | |
| Decrease | 37 | 5 | 1 |
| Maintain | 409 | 59 | 48 |
| Increase | 43 | 9 | 1 |

Table 11: The number of each event or option in each dataset split. Val. denotes the validation set.

# TRepLiNa: Layer-wise CKA+REPINA Alignment Improves Low-Resource Machine Translation in Aya-23 8B

**Toshiki Nakai**[1], **Ravi Kiran Chikkala**[1], **Lena Sophie Oberkircher**[1], **Nicholas Jennings**[1],
**Natalia Skachkova**[2], **Tatiana Anikina**[2], **Jesujoba O. Alabi**[1]

[1]Saarland University [2]German Research Center for Artificial Intelligence (DFKI)
`{toshiki3738,lenaoberkircher}@gmail.com`
`{rach00004@teams,s8nijenn@stud,jalabi@cs}.uni-saarland.de`

## Abstract

The 2025 Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo) Language Challenge addresses one of India's most pressing linguistic gaps: the shortage of resources for its diverse low-resource languages (LRLs). The challenge focuses on developing a translation model capable of translating between High resource languages (HRLs) (Hindi/English) and LRLs (Bhili, Mundari, Santali, and Gondi). In this study, we use the MMLoSo 2025 challenge dataset to investigate whether enforcing cross-lingual similarity in specific internal layers of a decoder-only multilingual large language model (LLM) can improve translation quality from LRLs to HRLs. Specifically, we combine Centered Kernel Alignment (CKA), a similarity metric that encourages representations of different languages to align with Representation Projection Invariance (REPINA), a regularization method that constrains parameter updates to remain close to the pretrained model, into a joint method, we call TRepLiNa (CKA + REPINA). Our results[1] show that aligning mid-level layers with TRepLiNa is a low-cost and practical way to improve LRL translation in data-scarce settings. We make our code and models public.

## 1 Introduction

Many multilingual LLMs share parameters across languages, yet transfer to low-resource languages (LRLs) often lags behind their performance on high-resource languages (HRLs) (Conneau et al., 2020; Zhang et al., 2020). Recent analysis of Aya-23 8B (Aryabumi et al., 2024), a multilingual decoder-only model, shows strong neuron overlap across related languages in the embedding layer, perhaps due to token overlap, but it exhibits a marked drop in overlap at intermediate and higher layers (Trinley et al., 2025). This suggests a simple hypothesis: *selectively increasing cross-lingual similarity where it is weakest (mid/high layers) may lead to better transfer for LRLs.* We focus only on the LRL→HRL translation, based on the intuition that models generally find it easier to understand a new language than to generate it (Lin et al., 2025). We operationalize this via a lightweight alignment loss between hidden representations of parallel sentences, which is applied at a chosen layer $\ell$. We use centered kernel alignment (CKA) (Kornblith et al., 2019), which can robustly compare representations across networks and layers, together with representation projection invariance (REPINA) (Razdaibiedina et al., 2023) to stabilize HRL features against representation drift. We perform experiments, using zero-shot (Zhao et al., 2023), few-shot (Karimi Mahabadi et al., 2022) and QLoRA-based fine-tuning (Zhang et al., 2023) on Aya-23 8B, using the MMLoSo benchmark (lrl, 2025) pairs, Hindi/English pivots as HRLs; Bhili (Indo-Aryan), Mundari (Austro-asiatic), Santali (Austro-asiatic) and Gondi (Dravidian) as LRLs.

Our work makes the following **contributions:**

- We present, to the best of our knowledge, the first systematic study of *layer-wise* alignment in a decoder-only LLM for low-resource machine translation (MT), comparing CKA and TRepLiNa (CKA+REPINA) across layers.
- We demonstrate that mid-layer alignment (roughly layers 10–15) is most effective, with TRepLiNa consistently favoring layer 15 in limited-data settings.
- We show improvements in the weighted composite score of BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), defined as ($0.6\times$BLEU$+0.4\times$ChrF) with TRepLiNa and provide guidelines on when and where alignment should be applied.

---

[1] https://github.com/konta3738/cka-repina-aya23

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 25–34
December 23, 2025 ©2025 Association for Computational Linguistics

## 2 Related Work

**Low-Resource Transfer Methods for Indic LRLs:** Alongside alignment-based methods, zero-shot and few-shot strategies have also been explored for Indic LRLs. Huidrom and Lepage (2020) show that a single multilingual Neural Machine Translation (NMT) model can translate between unseen Indian language pairs, with performance improving as small amounts of parallel data are added. Ghosal et al. (2025) address the problem of improving few-shot generation for Indic LRLs through prompt refinement for MT and other downstream generation tasks. Their findings highlight the importance of designing techniques that enhance low-resource performance. While they focus on input-level prompting, we complement this by aligning hidden representations across layers to improve transfer for Indic LRLs.

**Cross-lingual Alignment Methods:** Cross-lingual alignment has long been studied as a way to enhance transfer in multilingual models, particularly for LRLs (Hämmerl et al., 2024). Post-hoc cross-lingual alignment methods rotate representations after training, e.g., SVD/orthogonal Procrustes or projection-based removal of language-specific components, improving zero-shot transfer (Deb et al., 2023; Yang et al., 2021). Joint optimization injects alignment during training, e.g., cosine-similarity objectives on parallel sentences or contrastive InfoNCE setups (Wieting et al., 2019; Pan et al., 2021) while balancing negatives. CKA has emerged as a computationally attractive alternative to Canonical Correlation Analysis (CCA) (Hotelling, 1936) for comparing intermediate activations and for distillation/analysis (Dasgupta and Cohn, 2025). REPINA (Razdaibiedina et al., 2023) regularizes against representation collapse/drift. We apply these ideas to layer-wise alignment in Aya-23 8B for LRL MT.[2]

## 3 Data

In this research project, we use the MMLoSo shared task train dataset (lrl, 2025) for the experiments with roughly 20k sentence pairs per direction, splitting the dataset into 95% train and 5% development. The language pairs include Bhili↔Hindi, Mundari↔Hindi, Gondi↔Hindi (all in Devanagari script) and Santali↔English ( Santali in Ol Chiki script and English in Roman script). Our initial tokenization analysis of the data shows that Santali has higher tokenization fertility. It often requires a longer maximum sequence length (368) than Hindi/English (256), which can slightly reduce the tokenwise parallelism available to the alignment loss when sequences must be truncated to apply CKA.

## 4 Methodology and Experiments

In our experiments, we focus on Aya-23 8B, a strong openly available model with broad typological coverage and robust multilingual capabilities. The model is pretrained on 23 languages, including Hindi and English, but it does not cover Mundari, Bhili, Gondi, or Santali. We issue all prompts instructions in English.

### 4.1 Prompting

Here, we discuss the zero-shot and few-shot prompting methods that are used in the experiments.

**Zero-shot:** In zero-shot experiments, the model relies on its knowledge without any examples (Chikkala et al., 2025). We consider zero-shot as the baseline for the experiments. See Figure 3 for zero-shot prompt template in the Appendix.

**Few-shot:** In few-shot experiments, we use examples for each language pair from the train set as reference for the language model (Anikina et al., 2025). For each language pair, we use the first example from the training split of the provided data for one-shot, the first three for three-shot, and the first five for five-shot. See Figure 4 for few-shot prompt template in the Appendix.

### 4.2 TRepLiNa

This section describes the alignment objective of TRepLiNa. Figure 1 illustrates an overview of our proposed training method.

Given a parallel pair $(x^{(A)}, x^{(B)})$ from an LRL $A$ and a pivot HRL $B$, let $H_\ell^{(A)}, H_\ell^{(B)} \in \mathbb{R}^{T\times d}$ denote token wise hidden states at layer $\ell$ (sequence length $T$, width $d$) and $H_{\text{pre}\,\ell}^{(A)}$ be the hidden states obtained from the pretrained model (with an adapter disabled). We augment the MT loss (token-level cross entropy) $L_{\text{MT}}$ with (i) a CKA alignment between LRL/HRL representations and

---

[2] We focus on CKA here; exploring cosine/contrastive or newer similarity objectives (e.g., Listopad 2025) is left to future work.

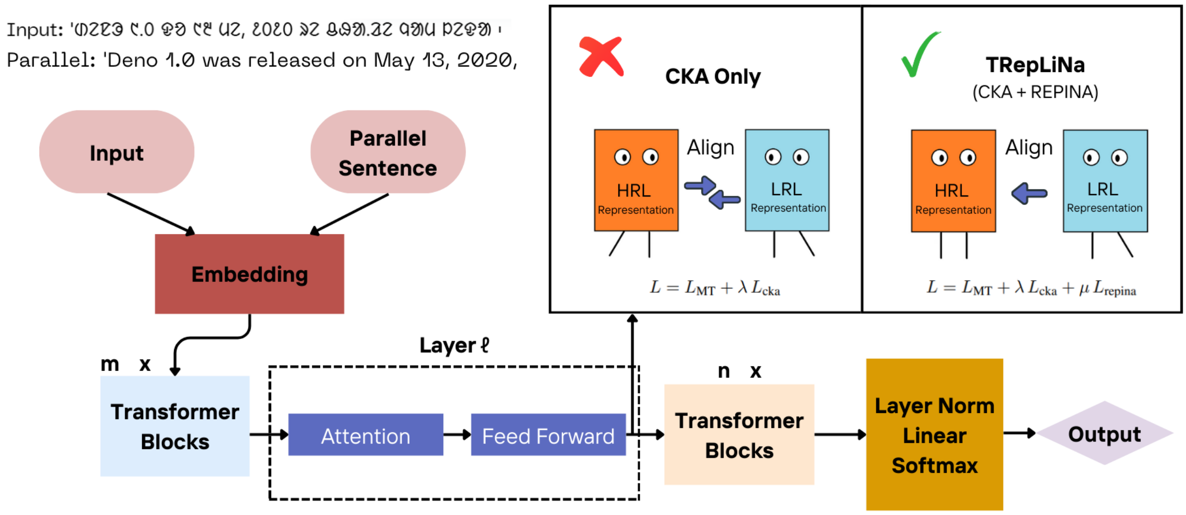


Figure 1: Proposed alignment architecture. Under **CKA-only**, both HRL and LRL representations drift toward each other, potentially distorting HRL features. By contrast, **TRepLiNa** constrains HRL representations while guiding LRL representations toward them, achieving targeted alignment without degrading HRL quality. Here, $m$ and $n$ denote the number of transformer blocks before and after the target alignment layer, respectively.

(ii) a REPINA anchoring term that resists drift of HRL features:

$$L = L_{\text{MT}} + \lambda\, L_{\text{CKA}} + \mu\, L_{\text{REPINA}} \tag{1}$$

with $\lambda, \mu > 0$. We use linear CKA on mean–centered features:

$$\begin{aligned} L_{\text{CKA}} &= 1 - \text{CKA}(H_\ell^{(A)}, H_\ell^{(B)}), \\ \text{CKA}(H_\ell^{(A)}, H_\ell^{(B)}) &= \frac{\|X^\top Y\|_F^2}{\sqrt{\|X^\top X\|_F^2}\,\sqrt{\|Y^\top Y\|_F^2}}. \end{aligned} \tag{2}$$

$F$ denotes Frobenius norm. $X$ and $Y$ represent the matrices after applying mean-centering on $H_\ell^{(A)}$ and $H_\ell^{(B)}$ respectively. For REPINA, we anchor HRL states to a stop-gradient identity mapping of a reference pass, i.e.,

$$L_{\text{REPINA}}(H_{\text{pre}\,\ell}^{(A)}, H_\ell^{(A)}) = \| H_{\text{pre}\,\ell}^{(A)} - \tilde{\phi}(H_\ell^{(A)}) \|_2^2, \tag{3}$$

Equivalently, $\tilde{\phi}(\cdot) = \text{sg}(\cdot)$; in our implementation this is the detached HRL hidden state at the same layer from the forward pass. CKA pulls $A$ toward $B$ at layer $\ell$, while REPINA stabilizes $B$. Unless noted, both terms are applied at a single layer $\ell$.

### 4.3 Experimental Design

**Step 1: Layer sweep (small data):** To make the sweep computationally tractable, we sample 1,000 parallel pairs and train for one epoch per direction (Mundari → Hindi, Santali → English). We sweep layers $\ell \in \{1, 2, 5, 10, 15, 20, 25, 30, 31, 32\}$ and evaluate **CKA-only** and **TRepLiNa (CKA+REPINA)** against two baselines **NoAlign** and **REPINA-only**. For **REPINA-only**, we fix $\ell = 15$ (the best layer observed under TRepLiNa) to isolate the marginal contribution of CKA. We set $\lambda = \mu = 0.05$, values that are large enough to reveal effects at this data scale, yet small enough to avoid the over-alignment; larger CKA weights (e.g., $\lambda = 0.3$) degraded MT performance in preliminary runs. The **NoAlign** (standard QLoRA finetuning) excludes both CKA and REPINA.

**Step 2: Longer training at the best layer:** Using the best layer from Step 1, we train for up to 5 epochs and track BLEU/ChrF on a 500-sample development set each epoch, comparing TRepLiNa vs. REPINA-only ($\lambda = 0.01$, $\mu = 0.05$).

## 5 Results and Analysis

Here, we analyze the results of zero-shot, few-shot, TRepLiNa, REPINA and NoAlign from Table 1

### 5.1 Step 1: Layer-Wise Trends

The result is discussed for 1k pairs and 1 epoch. For Mundari–Hindi, the weighted composite score across layers improves (see Figure 2). CKA peaks at layer 10, whereas TRepLiNa peaks at layer 15; the same tendency holds for Santali–English (see Appendix B.1).

**Interpretation:** CKA-only encourages both languages to meet in the middle; without a stabilizer, HRL features may drift, which can blunt gains in

| Language | Zeroshot | Few-shot (1) | Few-shot (3) | Few-shot (5) | TRepLiNa (Ours) | REPINA-only | NoAlign |
|---|---|---|---|---|---|---|---|
| Bhili→Hindi | 4.75 | 4.54 | 4.84 | 3.96 | 47.96 | **48.02** | 48.01 |
| Gondi→Hindi | 4.39 | 3.66 | 3.75 | 3.99 | **36.26** | 36.18 | 36.25 |
| Mundari→Hindi | 3.54 | 3.00 | 3.01 | 3.24 | **34.24** | 33.45 | 33.36 |
| Santali→English | 1.38 | 1.77 | 1.05 | 1.16 | **33.02** | 32.28 | 32.14 |

Table 1: Final translation scores across language pairs (0.6 × BLEU + 0.4 × ChrF). Best scores are in **bold.**

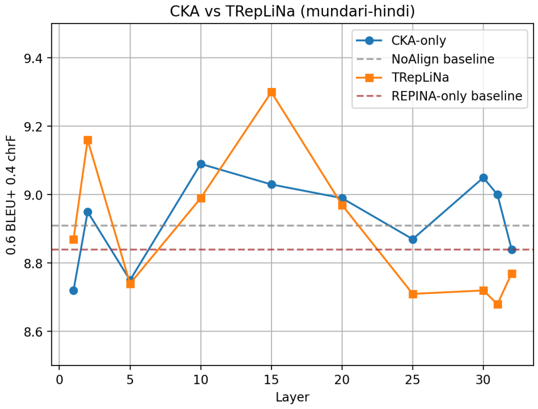


Figure 2: Comparison of (0.6× BLEU +0.4× ChrF) across layers for CKA, REPINA, NoAlign and TRepLiNa.

later layers. REPINA counteracts this, making mid-high layers (15) the sweet spot when pairing with CKA.

### 5.2 Step 2: Multi-Epoch Comparison at Selected Layer

**Setup:** Using the best alignment layer from Step 1 (typically a mid-layer around $\ell = 15$), we train for up to five epochs on the full split (≈20k pairs) and evaluate after each epoch on a 500-sample development set. Unless noted otherwise, we set $(\lambda, \mu) = (0.01, 0.05)$ for this longer run, i.e., a lower CKA weight than in Step 1 to avoid over-regularization at scale. We report the MMLoSo score (0.6×BLEU + 0.4×ChrF) and also track BLEU/ChrF separately (Appendix Table 2). Model selection uses the best development set checkpoint per direction.

### 5.3 Findings

**Gondi→Hindi:** TRepLiNa attains the highest performance score exceeding zero-shot performance. Few-shot(1) has the lowest score, the gap between the highest and lowest performance scores is 32.6.

**Mundari→Hindi:** TRepLiNa achieves the best score on development set outperforming zero-shot, while few-shot(1) has the lowest score, the difference between the best and the lowest performance score is 29.24.

**Santali→English:** TRepLiNa has the best performance score surpassing zero-shot, whereas Few-shot(3) has the lowest score. A difference of 31.97 exists between the best and worst performance scores. For comparison Billah et al. (2024) report a BLEU of 11.13 on their development set; our result (Appendix Table 2) is 25.24 BLEU, a +14.11 absolute and ≈2.27× relative improvement.

**Bhili→Hindi:** REPINA-only has the highest score, it could be because Bhili and Hindi are typologically close, a strong CKA weight can over-align and wash out beneficial language-specific features. However, our approach TRepLiNa performs better than zero-shot. Few-shot(5) has the lowest score, The highest score exceeds the lowest by 44.06.

**Takeaways:** (i) *Early vs. late epochs:* **NoAlign** shows stronger performance in the initial stages of training with 1k inputs (see Figure 2), whereas **REPINA-only** tends to surpass it when trained on larger datasets (20k). (ii) *Data scaling:* Larger datasets favor a lower CKA weight; we used $\lambda = 0.05$ for the 1k/1-epoch sweep and $\lambda = 0.01$ for 20k/5-epoch training. As cross-lingual representations become sufficiently aligned, excessive CKA pressure can erode language-specific cues. (iii) *Language proximity:* For related pairs (e.g., Bhili–Hindi), We recommend reducing $\lambda$; for more distant pairs, mid-layer TRepLiNa remains robust.

## 6 Conclusions

In this paper, we investigate layer-wise alignment as a simple and effective strategy for improving low-resource translation using Aya-23 8B on MMLoSo language pairs. We show that aligning representations at mid layers enhances performance on translation tasks between language pairs, and that coupling similarity (CKA) with stability (REPINA) in our proposed **TRepLiNa** method yields robust gains across data-scarce settings.

## Limitations

We do not explore other similarity objectives (cosine, contrastive InfoNCE) or recent proposals (Listopad, 2025); we use coefficients ($\lambda$, $\mu$) without scheduler/tuning; and this study does not include a thorough ablation study of the hyperparameters ($\lambda$, $\mu$). In our experiments, we have not explored chain of thought prompting techniques and different prompt templates. From the results Table 1, we observe that there is a reduced performance of TRepLiNa on Bhili→Hindi, where it underperforms the REPINA-only and NoAlign methods. These results indicate that our method may not generalize well to all language pairs. Santali tokenization sometimes requires longer sequences than 256, reducing token-wise overlap for alignment when truncation occurs. We do not evaluate human adequacy/fluency or domain transfer and qualitative analysis of the generated output by the models.

## Acknowledgments

We are grateful to our collaborators and mentors for valuable discussions on layer-wise analysis and low-resource MT. We especially thank Mina Abarico for designing the architecture diagram. This project was also supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005), and by Saarland University. We also thank the reviewers for their constructive comments. Any remaining errors are our own.

## References

2025. Multimodal models for low-resource contexts and social impact 2025. Kaggle Competition.

Tatiana Anikina, Ivan Vykopal, Sebastian Kula, Ravi Kiran Chikkala, Natalia Skachkova, Jing Yang, Veronika Solopova, Vera Schmitt, and Simon Ostermann. 2025. dfkinit2b at checkthat! 2025: Leveraging llms and ensemble of methods for multilingual claim normalization.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Syed Mohammed Mostaque Billah, Ateya Ahmed Subarna, Sudipta Nandi Sarna, Ahmad Shawkat Wasit, Anika Fariha, Asif Sushmit, and Arig Yousuf Sadeque. 2024. Towards santali linguistic inclusion: Building the first santali-to-english translation model using mt5 transformer and data augmentation. *Preprint*, arXiv:2411.19726.

Ravi Kiran Chikkala, Tatiana Anikina, Natalia Skachkova, Ivan Vykopal, Rodrigo Agerri, and Josef van Genabith. 2025. Automatic fact-checking in english and telugu. *ArXiv*, abs/2509.26415.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sayantan Dasgupta and Trevor Cohn. 2025. Improving language model distillation through hidden state matching. In *The Thirteenth International Conference on Learning Representations*.

Ujan Deb, Ridayesh Parab, and Preethi Jyothi. 2023. Zero-shot cross-lingual transfer with learned projections using unlabeled target-language data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 449–457, Toronto, Canada. Association for Computational Linguistics.

Soumya Suvra Ghosal, Soumyabrata Pal, Koyel Mukherjee, and Dinesh Manocha. 2025. PromptRefine: Enhancing few-shot performance on low-resource Indic languages with example selection from related example banks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 351–365, Albuquerque, New Mexico. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Rudali Huidrom and Yves Lepage. 2020. Zero-shot translation among Indian languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 47–54, Suzhou, China. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3638–3652, Dublin, Ireland. Association for Computational Linguistics.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Peiqin Lin, Marion Thaler, Daniela Goschala, Amir Hossein Kargaran, Yihong Liu, André F. T. Martins, and Hinrich Schütze. 2025. Construction-based reduction of translationese for low-resource languages: A pilot study on Bavarian. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 114–121, Vienna, Austria. Association for Computational Linguistics.

Aleksandr Listopad. 2025. Wave-based semantic memory with resonance-based retrieval: A phase-aware alternative to vector embedding stores. *Preprint*, arXiv:2509.09691.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Anastasia Razdaibiedina, Ashish Khetan, Zohar Karnin, Daniel Khashabi, and Vivek Madan. 2023. Representation projection invariance mitigates representation collapse. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14638–14664, Singapore. Association for Computational Linguistics.

Katharina Trinley, Toshiki Nakai, Tatiana Anikina, and Tanja Baeumel. 2025. What language(s) does aya-23 think in? how multilinguality affects internal language representations. *Preprint*, arXiv:2507.20279.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15590–15606, Toronto, Canada. Association for Computational Linguistics.

# A Appendix: Training and Implementation Details

## A.1 Codebase and Reproducibility

We provide a single-script trainer for QLoRA fine-tuning of Aya-23 with layer-wise alignment. Seeds are fixed for Python and PyTorch (CPU/GPU). All console/file logs are timestamped; training/eval logs are written via helper functions (`write_train_log`, `write_eval_log`). LoRA adapters are pushed to a Hugging Face repo using access tokens from environment variables.

## A.2 Model, Quantization, and LoRA

We load `CohereLabs/aya-23-8B` with 4-bit NF4 (BitsAndBytes) and `bf16` (or `fp16`). We enable `output_hidden_states` to obtain intermediate activations. LoRA is applied to standard projection modules `[q,k,v,o,gate,up,down]` with default $(r{=}16, \alpha{=}32, \text{dropout}{=}0.05)$. We use gradient checkpointing and `enable_input_require_grads()` to support k-bit training.

## A.3 Tokenization and Batching

We use a fast tokenizer; if the pad token is missing, EOS is used as PAD. Causal-LM inputs are left-padded; alignment-only passes are right-padded. Prompts follow: *"Translate to* {`lang_b_name`}*:*$\backslash n\{src\}\backslash n$*"*. Labels mask the prompt with $-100$. Max lengths are typically 256 (Santali uses 368). We pad to a multiple of 8 for tensor cores. Global batch size is 1 with gradient accumulation (default 16).

## A.4 Data Splits and Development Set

From a CSV with columns `src_col/tgt_col`, we create train/development set splits. If =<1k examples, development set =10%; otherwise ≈5% (capped 1k–2k). Development set evaluation uses up to 500 examples per epoch.

## A.5 Losses and Layer-wise Alignment (No Equations)

**Task loss:** We use a label-smoothed causal LM loss with $\epsilon = 0.1$ over valid target tokens.

**Alignment passes (procedure only):** For a parallel pair from LRL $A$ and HRL $B$, we:

1. Run *source-only* strings for both languages to collect hidden states at a chosen layer $\ell$.
2. Mask pads, align sequence lengths (truncate to maximum), flatten tokens across the batch, and mean-center features.
3. Compute a *similarity score* between $A$ and $B$ at layer $\ell$ and add its complement as an alignment penalty.

This is the same CKA objective introduced in the main text; we omit formulas here and refer the reader to the Methodology and Experiments section (Section 4).

**REPINA anchoring (procedure only):** Periodically (e.g., every two optimizer steps) we:

1. Disable adapters to obtain a *reference* HRL representation at layer $\ell$ on the same inputs.
2. Penalize the mean-squared deviation between current and reference HRL hidden states (stop-gradient on the reference).

This follows the REPINA scheme described in the main text; equations are intentionally omitted here.

**Combined objective:** Training minimizes task loss + similarity penalty + anchoring penalty with user-set coefficients (`--lambda_cka`, `--mu_repina`). Both terms are applied at a single chosen layer $\ell$.

## A.6 Optimization and Precision

We use PagedAdamW8bit (or AdamW) with $\beta = (0.9, 0.95)$, weight decay $0.01$, linear warmup (ratio default $0.05$), and LR in $[1 \times 10^{-4}, 2 \times 10^{-4}]$ (default $2 \times 10^{-4}$). Mixed precision uses `torch.amp.autocast` (`bf16/fp16`); for `fp16`, gradients use `GradScaler`. We clip global gradients to 1.0 for `bf16`. Gradients are zeroed with `set_to_none=True`. Optimizer steps occur every `grad_accum` micro-steps.

## A.7 Model and Training Defaults

Unless noted: max source/target 256 (Santali 368), LR $2 \times 10^{-4}$, warmup $5\%$, batch size 1, grad accumulation 16, and mixed precision. Layer $\ell$ is selected via sweeps; CKA and REPINA use the same $\ell$.

## A.8 BLEU and ChrF Results (Per Direction)

## Compute, Runtime, and Practical Notes

- **Hardware:** Experiments are ran on A100 40GB or H100 80GB (QLoRA fits comfort-

You are a translation assistant. Translate from Mundari {source} to Hindi {target} in Devanagari script.

Figure 3: Zero-shot prompt

You are a translation assistant. Translate from Hindi {source} to Mundari {target} in Devanagari script.

Example 1:

Hindi: बघइ प्रेतों पालनः- बघइ देवगण शिकार के समय मारे गये जानवरों के माँस और बाघ के द्वारा मारे गये आदमी के माँस से अपने पाल रहे हैं।,

Mundari: बघइअ बोंगाकोअः अनसुल- बघइअ बोंगाको सेनदेरा तेको गोएःकेद् बिर जिलु ओड़ोः कुला गोएःकि होड़ो जिलुतेको असुलनतना।

.

.

.

Example 5:

Hindi: यह....

Mundari: वे कौन....

Figure 4: Few-shot prompt

ably); BF16 preferred when available. Training took approximately 30 hours on 1 A100 40GB, and 16 hours on 1 H100 80GB.

- **Stability:** For typologically close pairs (e.g., Bhili–Hindi), reduce the similarity weight over epochs to avoid over-alignment.
- **Layer indexing:** Hidden state tuple index 0 corresponds to the embedding output; a user layer $\ell$ refers to the 1-based transformer block output.

# B Appendix B: Complementary Results

## B.1 Step-1: Layer Sweep on Santali→English

With only 1,000 training pairs and a single epoch, *anchoring* from REPINA can transiently conflict with task updates: large anchoring ($\mu$) tends to pull parameters back toward the reference HRL representation, partially canceling early task learning. Empirically, $\lambda$=0.05, $\mu$=0.05 underperforms **CKA**-only, but reducing anchoring to $\mu$=0.01 makes **TRepLiNa** outperform **CKA**-only. Performance peaks at $\ell$=15, suggesting a mid-layer is most effective for aligning Santali to English in

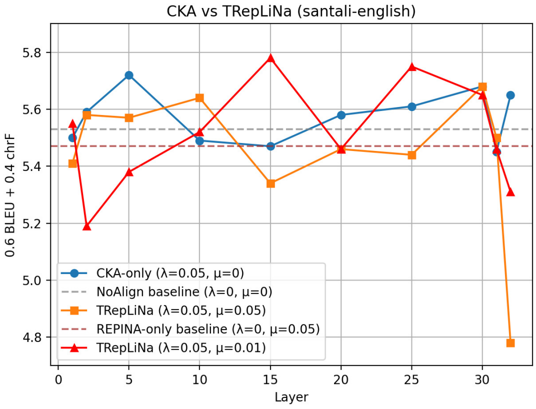


Figure 5: Comparison of (0.6 × BLEU + 0.4 × ChrF) across layers for **CKA** and **TRepLiNa** on Santali→English (1k rows, 1 epoch). Dashed lines indicate each method's baseline.

| Language | Zeroshot | Few-shot (1) | Few-shot (3) | Few-shot (5) | TRepLiNa (Ours) | REPINA-only | NoAlign |
|---|---|---|---|---|---|---|---|
| Bhili→Hindi | 0.88 | 0.64 | 0.93 | 0.35 | 40.15 | **40.26** | 40.13 |
| Gondi→Hindi | 0.37 | 0.12 | 0.30 | 0.56 | **28.71** | 28.44 | 28.64 |
| Mundari→Hindi | 0.14 | 0.06 | 0.04 | 0.08 | **25.94** | 25.08 | 24.93 |
| Santali→English | 0.04 | 0.04 | 0.03 | 0.05 | **25.24** | 24.64 | 24.26 |

Table 2: Final translation scores across language pairs (BLEU). Best scores are in **bold.**

| Language | Zeroshot | Few-shot (1) | Few-shot (3) | Few-shot (5) | TRepLiNa (Ours) | REPINA-only | NoAlign |
|---|---|---|---|---|---|---|---|
| Bhili→Hindi | 10.57 | 10.40 | 10.72 | 9.38 | 59.67 | 59.65 | **59.84** |
| Gondi→Hindi | 10.42 | 8.97 | 8.93 | 9.12 | 47.58 | **47.78** | 47.67 |
| Mundari→Hindi | 8.66 | 7.43 | 7.48 | 7.98 | **46.68** | 46.02 | 46.00 |
| Santali→English | 3.40 | 4.39 | 2.60 | 2.83 | **44.68** | 43.74 | 43.96 |

Table 3: Final translation scores across language pairs (ChrF). Best scores are in **bold.**

this small-data setting. **Practical note:** for low data/short training, prefer moderate CKA ($\lambda \approx 0.05$) with lighter anchoring ($\mu \approx 0.01$) and sweep mid-layers (e.g., 10–20).

### B.2 BLEU Table: Summary and Takeaways

Table 2 compares final BLEU across settings. On **Mundari→Hindi** and **Santali→English**, **TRepLiNa (CKA+REPINA)** achieves the best scores, outperforming both *REPINA-only* and *NoAlign*. For **Bhili→Hindi**, *REPINA-only* narrowly leads. Few-shot and zero-shot remain far below alignment-based methods, indicating that explicit layer-wise alignment is crucial in the low-resource regime.

### B.3 ChrF Table: Summary and Takeaways

Table 3 shows the same comparison in ChrF. The pattern largely mirrors BLEU: **TRepLiNa** tops **Mundari→Hindi** and **Santali→English**, while *NoAlign* is slightly best on **Bhili→Hindi**. Despite small differences between top systems on Bhili→Hindi, both metrics agree that alignment generally helps, especially for the more distant pairs. Overall, ChrF confirms the BLEU trends and supports the utility of combining CKA with REPINA.

## C Appendix C: Future Directions

**Scope:** We did not explore HRL → LRL directions in the main paper due to the asymmetric computational profile of the task and the cost of fine-tuning Aya-23 8B. Here we provide a preliminary *Step-1* layer sweep on Hindi→Mundari (1k pairs, 1 epoch; $\lambda = 0.05$, $\mu = 0.05$).

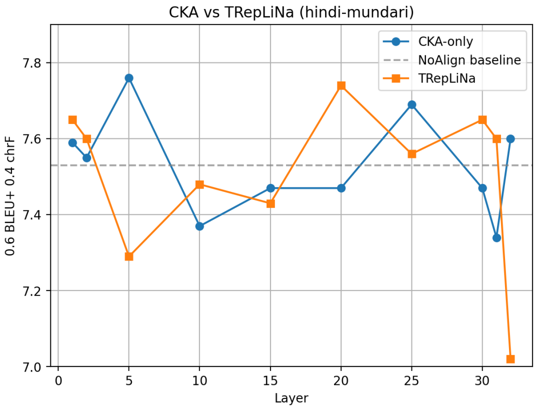


Figure 6: Layer sweep on Hindi→Mundari (1k pairs, 1 epoch). We plot 0.6 × BLEU + 0.4 × ChrF for **CKA-only** and **TRepLiNa**; dashed lines denote each method's NOALIGN baseline. CKA-only peaks at $\ell$=10, TRepLiNa at $\ell$=20.

**Setup and metrics:** We compare **CKA-only** and **TRepLiNa** against the **NoAlign** baseline across layers, using the combined score $0.6 \times \text{BLEU} + 0.4 \times \text{ChrF}$ (Figure. 6).

**Observations:** **(i)** CKA-only peaks at layer 10 and TRepLiNa peaks at layer 20; both outperform NOALIGN. **(ii)** With $\mu = 0.05$ and such a small regime (1k/1 epoch), REPINA can over-regularize, likely dampening short-term task learning. This suggests TRepLiNa may be more competitive under larger budgets (e.g., 20k/5 epochs), where the auxiliary signal has time to synergize with the task objective.

**Layer asymmetry:** For LRL → HRL, we observed peaks around layers 10–15 for TRepLiNa, whereas HRL → LRL peaks later (layer 20). One plausible explanation is that Aya-23 8B has limited pretrained support for LRL tokens and structures. When the *output* is an LRL (e.g., Mundari), later layers must adapt themselves to generate unseen languages; when the *input* is an LRL, earlier layers need to map LRL signals into language-agnostic features. We leave a rigorous verification of this hypothesis to future work.

Future work may extend this approach to encoder–decoder or speech–text models, and explore adaptive scheduling strategies for alignment strength in truly low-data scenarios.

# MemeGuard: Transformer-Based Fusion for Multimodal Propaganda Detection in Low-Resource Social Media Memes

**Md. Mohiuddin, Kawsar Ahmed**
**Shawly Ahsan** and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
`{u1904103, u1804017}@student.cuet.ac.bd`
`shawly.ahsan.bd@gmail.com, moshiul_240@cuet.ac.bd`

## Abstract

Memes are now a common means of communication on social media. Their humor and short format help messages spread quickly and easily. Propagandistic memes use both words and images to influence opinions and behaviors, often appealing to emotions or ideologies. While propaganda detection has been well-studied in high-resource languages (HRLs), there has been a limited focus on low-resource languages (LRLs), such as Bengali. In this study, we introduce **MemeGuard**, a new dataset of 3,745 memes for detecting propaganda in Bengali. We tested more than 45 different methods, including both single and combined approaches with fusion. For text, BanglaBERT-1 achieved the best macro F1 score of 80.34%, whereas the CLIP vision transformer scored 78.94% for images. The proposed multimodal model, which combines BanglaBERT-2 and CLIP via Adaptive Modality Fusion, achieved the highest macro-F1 score of 85.36%. This work establishes a strong baseline and offers valuable insights for future research in Bengali multimodal content analysis.

## 1 Introduction

Digital platforms have transformed human interaction by altering the ways individuals connect, share information, and express themselves. The proliferation of the Internet and Web 2.0 applications has fostered large, dynamic online communities, enabling rapid and accessible communication. Although digital openness offers significant advantages, it also accelerates the dissemination of misleading information, manipulative influence, and harmful narratives. On social media, memes function as a prominent and efficient communication medium, integrating concise text with impactful visuals to transmit messages (Zhong and Baghel, 2024). Certain memes are intentionally designed to manipulate audiences, influence opinions, and promote bias. These propagandistic memes advance specific political, religious, cultural, or ideological agendas by exploiting emotional responses or distributing misinformation (Cheng, 2025). As the influence of memes grows, identifying and analyzing propagandistic content has become essential. Because meme interpretation depends on both textual and visual elements, multimodal analysis is required for accurate detection and classification. While substantial research has addressed propaganda detection using text-based or multimodal methods in HRLs, LRLs such as Bengali remain underexplored. A comprehensive multimodal analysis for Bengali content is currently unavailable (Hossain et al., 2025). Detecting propagandistic memes in Bengali is necessary to support the dissemination of accurate information. However, no prior research has investigated propaganda detection in Bengali memes, leaving a significant research gap despite the growing prevalence of such content.

Building an automated system to detect propagandistic memes in Bengali presents several challenges. One major issue is the lack of a public dataset and the difficulty of extracting Bengali text from images, since there is no standard OCR tool for the language. Labelling memes by hand is also tricky because propaganda can be interpreted differently by different people. Memes often combine images and text, and the same image with different text can convey different meanings, adding complexity. Other problems include short text, discrepancies between the image and text, and the need to integrate both types of information. To address these problems, this study introduces a dataset of 3,745 Bengali-language memes for detecting propaganda. The study also proposes a transformer-based model that utilizes BanglaBERT-2 and CLIP, with an adaptive fusion of text and image features, to identify propaganda in memes more effectively. The main contributions of this work are:

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 35–47
December 23, 2025 ©2025 Association for Computational Linguistics

- Developed **MemeGuard**, a multimodal dataset containing 3,745 memes, labelling propagandistic and non-propagandistic.
- Introduced a multimodal framework that combines textual and visual features using a late fusion strategy, where BanglaBERT-2 and CLIP models are employed with adaptive modality fusion to detect propaganda in memes effectively.

## 2 Related Work

Several studies have been conducted in various languages to detect propaganda in memes, including text, images, and multimodal content. This section provides a brief review of past studies on detecting memes, specifically propagandistic memes, across unimodal (e.g., text and image) and multimodal content.

### 2.1 Unimodal-based Propaganda Detection

Text-based propaganda detection has progressed considerably. Early work applied ML/DL with word embeddings. Noman et al. (2024) used a BiLSTM-CRF model for semantic web-based propaganda text, reporting F1 scores of 0.61 on multilingual and 0.688 on monolingual news data. Lichouri et al. (2023) examined disinformation detection using surface and morphological preprocessing, FastText vectors, and weighted TF-IDF fusion, obtaining a 77.60% F1-micro with LSVC, though effectiveness remained limited. A three-stage framework (Sourati et al., 2023) targeted logical fallacies in manipulative text. Building on such work, transformers and LLMs have been widely used for text classification, including propaganda detection. Ojo et al. (2023) conducted binary detection of persuasion strategies in Arabic news and tweets, achieving 64.00% F1 with XLM-RoBERTa. Horák et al. (2024) reported 92.26% F1 using XLM-RoBERTa Large for newspaper texts. Salman et al. (2023) found strong performance for code-switched English–Roman Urdu social media text using XLM-RoBERTa (Roman Urdu) and GPT-3.5-Turbo. Hasanain et al. (2024a) noted AraBERT outperforming GPT-4 for news articles, while Piña-García (2025) applied LLaMA 3.2 to political propaganda on Twitter.

In comparison, propaganda analysis using visuals alone has received far less attention than text-based approaches. Hs et al. (2021) used a DL-based ResNet-50 model and achieved 48.00% F1. More recently, Wang and Chen (2025) introduced a hybrid method for image-based propaganda detection. Koutlis et al. (2023) proposed Visual Part Utilization (VPU) with a ViT, reaching 94.98% accuracy but still excluding text. However, unimodal text- or image-only approaches fail to capture subtle context and often struggle with patriarchal content, underscoring the need for models that handle the complexities of multimedia content.

### 2.2 Multimodal-based Propaganda Detection

In addition to unimodal analysis, several multimodal approaches have been explored. Zaytoon et al. (2024) combined Bloomz-1b1 and ResNet101 with concatenation fusion for meme propaganda detection, achieving 80.51% F1-macro, though results were limited by an imbalanced Arabic dataset. Mahmoud and Nakov (2024) used VLM-generated descriptions with MPNet and CLIP-ViT for propaganda detection and multilabel classification, reporting 66.67% F1-macro but facing severe imbalance issues. Dimitrov et al. (2021) introduced a 950-meme corpus with 22 strategies, where VisualBERT COCO achieved 48.34% F1-micro, constrained by the small, imbalanced dataset. Alam et al. (2024a) created a 6,000-meme Arabic corpus with four classes, yielding weighted F1 scores of 69.00% (Qarib), 67.30% (ResNet50), and 65.90% (ConvNeXt, AraBERT, SVM), again limited by class imbalance. Qu et al. (2022) released Disinfomeme, a 1,170-meme dataset labelled as Disinfo or Non-Disinfo, where VisualBERT COCO achieved 53.3% on the BLM subset and 30.60% on the Veganism subset. Overall, these studies highlight the complexity of the task and suggest that improved fusion-based approaches may offer performance gains.

### 2.3 Multimodal Content Detection in Bengali

In contrast to other languages, propaganda detection using multimodal content in Bengali remains at a rudimentary stage. Existing multimodal studies using DL and transformer-based models have mainly addressed fake news detection (FAR, 2025), hate speech detection (Hossain et al., 2022), emotion classification (Rahman et al., 2025; Das et al., 2024), aggression detection (Hasan et al., 2025), and commercial content detection (Shanto et al., 2025). Multimodal content such as memes has also been used to detect aggressiveness (Alam et al., 2024b) and for sentiment analysis (Ahammad et al., 2025). Work leveraging LLMs for multimodal classification is similarly limited. Hasan et al. (2024) examined LLMs with zero- and few-shot techniques for Bengali sentiment analysis and observed inferior performance with GPT-4. Building on this, Hossain et al. (2025) investigated VLMs, LLMs, MLMs, and vision transformers for multimodal text classification, showing that integrating pre-trained vision transformers for visual encoding and MLMs for textual encoding through fusion produced the best results.

Most existing studies in Bengali have primarily focused on detecting fake news, sentiment, aggression, and emotion, using mainly memes and text-image pairs. However, the detection of propaganda remains unexplored, and, to the best of our knowledge, there is currently no publicly available multimodal dataset specifically designed for propaganda detection, nor has a comprehensive multimodal analysis been conducted in this context. To address this gap, this study examines various multimodal techniques for the task on a newly developed dataset.

## 3 Dataset Development: MemeGuard

Our research reveals that no dataset currently exists for identifying multimodal propaganda in Bengali memes. To fill this gap, we developed **MemeGuard**, a multimodal dataset of 3,745 samples. This section describes the dataset's development process and key statistics.

### 3.1 Data Accumulation

The dataset was compiled by collecting Bengali memes over a five-month period from two sources: Facebook and a curated archive (Alam et al., 2024b). Facebook, a lively hub for Bengali user-created memes, provided dynamic content that captured local humour and cultural nuances. The archive offered structured collections that reflected diverse styles, in which textual variations could alter meaning. Keywords such as Bengali Memes, Bengali Funny Memes, Propaganda Memes, Bengali Celebrity Memes, Bengali Offensive Memes, and Bengali Political Memes were used to search groups and collect memes.

A total of 3,745 memes were collected, comprising 1,501 (40.1%) from Facebook and 2,244 (59.9%) from the curated archive (Fig. 1).

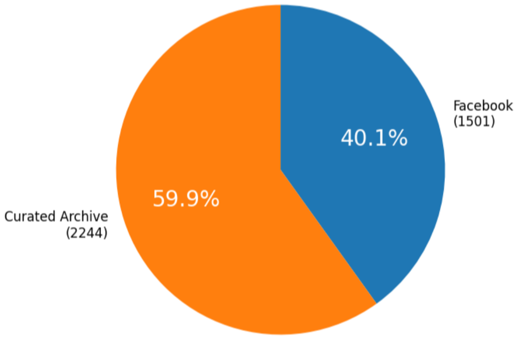


Figure 1: Distribution of data collection sources

Only memes with Bengali captions were included, while excluding the following memes: (i) unimodal memes (text-only or image-only), (ii) memes containing unreadable text or very low quality image, and (iii) already existing memes. After that, the text was manually extracted from the browser using Google Lens, as Bengali lacks a reliable OCR system. Only relevant content was captured, excluding extraneous elements such as group names or creator identifiers. Finally, the extracted texts from memes were passed to annotators for manual labelling to ensure a rich dataset.

### 3.2 Dataset Annotation

The MemeGuard dataset is designed for binary classification, categorizing memes into two distinct categories: Non-Propagandistic (Non-Prop) and Propagandistic (Prop). We follow the propaganda techniques proposed by Dimitrov et al., 2021 to define these categories in a simple, engaging, and precise way.

- **Non-Prop**: These memes are playful and neutral, designed to entertain without advancing any agenda. They often use humor, everyday scenarios, or light satire to foster connection through shared amusement.

- **Prop**: These memes seek to influence opinions or actions toward a specific goal. They promote political, ideological, or social agendas, often using emotionally charged or misleading content.

Guidelines for annotators are crucial for ensuring high-quality datasets. To assist annotators, we identified key characteristics as questions (see Appendix A), drawn from established propaganda techniques (Dimitrov et al., 2021), which are critical for distinguishing their manipulative nature. A team of five members conducted the manual annotation: four early-career NLP researchers (three graduate students and one research assistant) and one senior NLP expert with 23 years of experience. The early-career annotators had 1–2.5 years of NLP experience, with 2 of them having prior annotation experience. Their ages ranged from 24 to 26, while the expert was 48.

A meme is considered propagandistic if it meets one or more characteristics defined in Appendix A. In the first stage, the three graduate annotators independently label the memes. Majority voting is applied to their labels to create the initial dataset. The RA annotator reviews this preliminary dataset. If the RA finds inconsistencies in labeling, the cases are discussed with the Expert Annotator to reach a final decision. This process produces the finalized dataset. We then calculated Cohen's kappa coefficient to measure inter-rater agreement (Cohen, 1960), with an average kappa value of 0.84. This shows nearly perfect agreement on the kappa scale, as shown in Table 8 (Appendix B), highlighting the robustness and dependability of the annotations for Bengali propagandistic meme detection.

### 3.3 Dataset Statistics

The MemeGuard dataset comprises 3,745 Bengali memes, with 865 labeled as *Prop* and 2,880 labeled as *Non-Prop*. The text contains 1,881 unique words, reflecting the linguistic diversity of Bengali memes. To support model training and evaluation, the dataset was stratified into 70% (2,621 memes) for training, 15% (562 memes) for validation, and 15% (562 memes) for testing, ensuring proportional representation of the 865 propagandistic and 2,880 non-propagandistic memes in each subset, as shown in Table 1. Specifically, the training set contains 605 propagandistic and 2,016 non-propagandistic memes, while the validation and testing sets each include 130 propagandistic and 432 non-propagandistic memes. Textually, memes average around 14 words per sample, with sentence lengths ranging from 2 to 66 words, and a total vocabulary exceeding 11,000 unique words. Visually, all images were resized to a uniform resolution of 224 × 224 pixels and stored in either PNG or JPEG format, with an overall average size of approximately 118 KB. This standardized and stratified setup ensures consistency, class balance, and reproducibility for robust model training and evaluation.

Figure 2 presents the frequency distribution of text lengths across the dataset, revealing that most memes contain between 5 and 20 words.

| Class | Train | Validation | Test | *Total* |
|---|---|---|---|---|
| Prop | 605 | 130 | 130 | 865 |
| Non-Prop | 2016 | 432 | 432 | 2880 |
| **Total** | 2621 | 562 | 562 | 3745 |
| $T_S$ | 2621 | 562 | 562 | 3745 |
| $T_W$ | 36340 | 8216 | 7834 | 52390 |
| $T_{UW}$ | 9145 | 3444 | 3287 | 11308 |
| $L_{min}$ | 2 | 3 | 2 | - |
| $L_{max}$ | 55 | 66 | 47 | - |
| $L_{avg.}$ | 13.93 | 14.73 | 13.96 | - |
| $I_T$ | 2621 | 562 | 562 | 3745 |
| $I_{avg.}$ (KB) | 117.05 | 124.42 | 114.97 | 117.84 |
| $I_R$ (px) | **224 × 224** | | | |
| $I_{format}$ | **PNG or JPEG** | | | |

Table 1: Distribution of data across train, validation, and test sets. The symbols $T_S$, $T_W$, $T_{UW}$ denote the total sentences, total words, and total unique words.

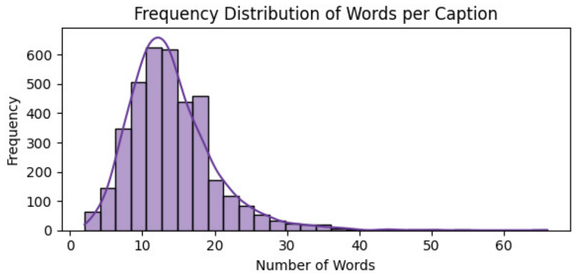


Figure 2: Frequency distribution of words per text

## 4 Methodology

A meme combines visual and textual elements, requiring parallel processing of both. This work examines deep learning and transformer-based models to extract and integrate these features. Feature-level fusion is then used to classify memes as propagandistic or non-propagandistic.

### 4.1 Data Preprocessing

Preprocessing standardizes inputs across modalities, thereby optimizing both learning and inference performance. Textual content preprocessing prepares data for classification models. Raw text is cleaned by removing stopwords, punctuation, and other unnecessary characters, reducing noise and improving feature extraction. The processed text is converted into dense vectors using embeddings like GloVe and FastText, or tokenized into IDs and attention masks with the Hugging Face tokenizer for transformer models such as BanglaBERT, MuRIL, and XLM-R. Padding adds extra tokens for uniform input lengths. Text normalization standardizes the format to ensure compatibility with pretrained vocabularies, especially for language-specific models like BanglaBERT.

For the visual modality, each image is resized to 224 × 224 pixels with three colour channels and converted to a tensor (a multidimensional array) using PyTorch, enabling GPU acceleration and batch processing. Normalization is performed using mean and standard deviation values from the ImageNet dataset, ensuring compatibility with the input requirements of standard image models such as VGG16, ResNet50, Swin Transformer, and Vision Transformer (ViT).

### 4.2 Unimodal Baselines

- **Text Modality:** Propagandistic meme detection explored various unimodal baselines that leverage deep learning and transformer-based architectures. For the text modality, CNN, BiLSTM, and CNN-BiLSTM hybrids were employed for binary classification, utilizing 300-dimensional GloVe and FastText embeddings. Training utilized the Adam optimizer, binary cross-entropy loss, and callbacks such as EarlyStopping and ReduceLROnPlateau. Transformer models included BanglaBERT-1 (Sarker, 2021), BanglaBERT-2 (Bhattacharjee et al., 2021), mBERT (Devlin et al., 2019), MuRIL (Khanuja et al., 2021), IndicBERT (Kakwani et al., 2020), Bangla-Electra (NLP, 2024), and XLM-R (NLP, 2024). These pretrained Hugging Face models were fine-tuned on the developed corpus. Appendix C lists the tuned hyperparameters utilized for DL and transformer models for text modality.

- **Visual Modality:** The visual modality was addressed by sequentially applying pretrained convolutional neural networks (VGG16, VGG19, ResNet50, EfficientNet-B0, EfficientNet-B3 (Tan and Le, 2019)) and transformer-based models (ViT (Dosovitskiy et al., 2020), Swin (Liu et al., 2021), BEiT (Bao et al., 2021), DeiT (Touvron et al., 2021), ConvNeXT (Liu et al., 2022), CLIP (Radford et al., 2021)) for image feature extraction, each with varied fine-tuning strategies. Specifically, VGG19 (Simonyan and Zisserman, 2014) was fine-tuned on features [:15] using a custom classifier with linear layers, batch normalization, ReLU, and dropout. VGG16 was tuned up to features [:20] with fully connected layers, ReLU, and dropout. EfficientNetB3 unfroze the last 30 layers and employed global average pooling, batch normalization, dropout, and dense layers (ReLU, sigmoid). ResNet50 unfroze the previous 20 layers and used dropout and dense layers. EfficientNetB0 unfroze layers from 100 onward, with global average pooling, dropout, dense layers (ReLU), and sigmoid activation. For transformer-based models, pretrained versions were obtained from the Hugging Face collection and fine-tuned on the developed dataset. All models were trained with binary cross-entropy loss and class weights to address imbalance. Appendix C presents the various tuned hyperparameters used to create multiple visual models.

### 4.3 Multimodal Baselines

This work explores 16 multimodal baselines generated by combining the top-performing four textual (BanglaBERT-1, BanglaBERT-2, MuRIL, XLM-R) and four visual models (CLIP, BEiT, ViT, Swin), which merge their respective complementary strengths with various hyperparameters (Table 9 in Appendix C). The models were combined using feature-level fusion, where the [CLS] features from the textual and visual models were fused before generating logits, enabling effective integration of both modalities and enhancing the overall performance of the multimodal system.

#### 4.3.1 Proposed Methodology

The proposed architecture for propagandistic meme detection in Bengali integrates two pre-trained models: BanglaBERT-2 for processing textual data and CLIP for analyzing image data. These models are combined using a late fusion approach, as illustrated in Figure 3, to leverage information from both modalities effectively. BanglaBERT-2 and CLIP models are fine-tuned on the developed dataset with manually tuned hyperparameters; both models are obtained from Hugging Face. To ensure consistency across modalities, both models were fine-tuned with carefully selected configurations derived through extensive manual experimentation. Let $V_{\text{logits}} \in \mathbb{R}^C$ and $T_{\text{logits}} \in \mathbb{R}^C$ denote the class logits produced by the visual and textual models, respectively, where $C$ is the number of classes. The fusion procedure can be described as follows.

To calibrate the sharpness of each distribution, we apply temperature scaling with a learnable pa-

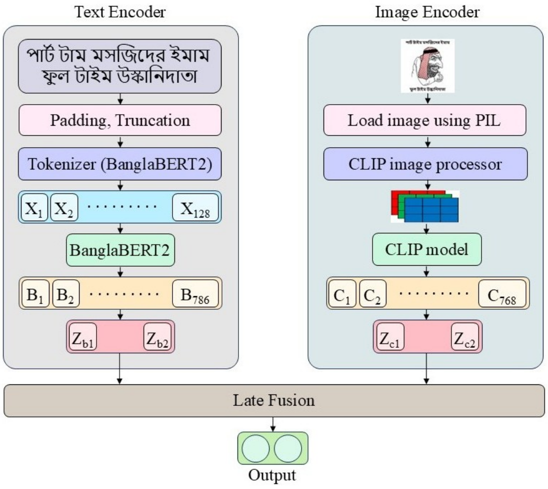


Figure 3: Architecture of the proposed multimodal approach for propaganda detection from memes.

rameter $\tau > 0$, with an initial value of 1.0 (Eq. 1).

$$V_{\text{scaled}} = \frac{V_{\text{logits}}}{\tau}, \qquad T_{\text{scaled}} = \frac{T_{\text{logits}}}{\tau}. \tag{1}$$

To adaptively control the relative contribution of each modality, we introduce learnable parameters $w_1, w_2 \in \mathbb{R}$ and compute modality weights via a softmax operation instead of giving equal weight as illustrated in Eq. 2.

$$\begin{aligned} [\alpha_1, \alpha_2] &= \text{softmax}([w_1, w_2]), \\ \alpha_1 + \alpha_2 &= 1, \quad \alpha_i \geq 0. \end{aligned} \tag{2}$$

The fused logits are obtained as a convex combination of the scaled logits (Eq. 3.

$$\text{Final\_Logits} = \alpha_1 \cdot V_{\text{scaled}} + \alpha_2 \cdot T_{\text{scaled}} \tag{3}$$

Finally, the predicted label is chosen as the class with the largest value in the fused logits, as shown in Eq. 4.

$$\hat{y} = \arg \max_{c \in \{1,\ldots,C\}} \text{Final\_Logits}[c] \tag{4}$$

## 5 Experiments

The proposed framework was implemented and tested on a Kaggle GPU instance with a Tesla P100 GPU, 30 GB of RAM, and a Linux operating system supporting CUDA 11.8 and cuDNN. Python was used for development, utilizing PyTorch 2.1.0 and Hugging Face Transformers 4.35.2 for deep learning, as well as Pandas 2.0.3 and NumPy 1.24.3 for data processing. Experiments were run in Kaggle's Jupyter Notebook environment. Model performance was assessed using macro-F1, weighted-F1, and geometric mean (G-mean). All code and data are publicly available at https://github.com/MohiuddinPrantiq/MemeGuard-MultimodalPropagandaDetection.

### 5.1 Results and Analysis

Although the primary criterion for selecting the top-performing model was the macro F1-score (M-F1), which is suitable for imbalanced datasets, additional metrics, such as the weighted F1 score (W-F1) and G-Mean (G), were used for a comprehensive performance comparison.

### 5.1.1 Performance of Unimodal Baselines

Table 2 presents the performance of textual baselines. CNN+BiLSTM with FastText embeddings achieved superior performance (79.36&) over other DL models. Notably, Bangla BERT-1 achieved the highest M-F1 score of 80.34%, outperforming all other textual models tested in this study.

| Model | M-F1 | W-F1 | G |
|---|---|---|---|
| CNN+GloVe | 0.7069 | 0.7823 | 0.7212 |
| CNN+ FastText | 0.7069 | 0.7823 | 0.7212 |
| BiLSTM+GloVe | 0.6900 | 0.7600 | 0.7246 |
| BiLSTM+FastText | 0.7800 | 0.8400 | 0.7960 |
| CNN+BiLSTN+ GloVe | 0.7029 | 0.7717 | 0.7397 |
| **CNN+BiLSTN+FastText** | **0.7936** | 0.8491 | 0.8080 |
| MuRIL | 0.8011 | 0.8563 | 0.8070 |
| BanglaBERT-2 | 0.7970 | 0.8549 | 0.7939 |
| IndicBERT | 0.7479 | 0.8159 | 0.7560 |
| m-BERT | 0.7879 | 0.8516 | 0.7666 |
| XLM-R | 0.8002 | 0.8550 | 0.8091 |
| Bangla-Electra | 0.7111 | 0.7793 | 0.7452 |
| **BanglaBERT-1** | **0.8034** | 0.8624 | 0.7826 |

Table 2: Performance of textual models.

Table 3 illustrates the performance of visual baselines. Among DL models, VGG19 and ResNet50 demonstrated strong performance, with VGG19 achieving the higher M-F1 score of 76.27%. However, CLIP achieved the highest macro F1-scores of 78.94%, demonstrating the superior representational capacity of transformer architectures in visual feature extraction for propagandistic meme detection.

### 5.1.2 Performance of Multimodal Models

Multimodal analysis, which utilizes both textual and visual modalities, enhanced performance across a feature-level fusion strategy, enabling effective integration and improving the overall performance of the multimodal system. As shown in Table 4, among the top four performing models per modality, **BanglaBERT-2+CLIP** achieved

| Model | M-F1 | W-F1 | G |
|---|---|---|---|
| VGG16 | 0.5734 | 0.6219 | 0.6562 |
| **VGG19** | **0.7627** | 0.8438 | 0.6939 |
| ResNet50 | 0.7500 | 0.8200 | 0.7386 |
| EfficientNet-B0 | 0.4600 | 0.6800 | 0.1723 |
| EfficientNet-B3 | 0.5500 | 0.7000 | 0.4473 |
| ViT | 0.7749 | 0.8400 | 0.7491 |
| BEiT | 0.7861 | 0.8534 | 0.7460 |
| **CLIP** | **0.7894** | 0.8587 | 0.7302 |
| ConvNeXT | 0.7384 | 0.8267 | 0.6642 |
| DeiT | 0.7643 | 0.8400 | 0.7367 |
| Swin | 0.7649 | 0.8400 | 0.7190 |

Table 3: Performance of visual models.

the highest M-F1 score of 82.83%. BanglaBERT-1+CLIP followed closely with an M-F1 score of 82.62%. Notably, CLIP consistently contributed to top-performing results across text models, highlighting its strong visual representation capabilities. The MuRIL – SWIN pair also demonstrated competitive performance, with a macro F1-score of 80.99%.

| Text | Image | M-F1 | W-F1 | G |
|---|---|---|---|---|
| BanglaBERT-1 | CLIP | 0.8262 | 0.8785 | 0.8056 |
| | BEiT | 0.7970 | 0.8584 | 0.7733 |
| | ViT | 0.7837 | 0.8549 | 0.7243 |
| | Swin | 0.7979 | 0.8636 | 0.7444 |
| MuRIL | CLIP | 0.7673 | 0.8336 | 0.7614 |
| | BEiT | 0.7586 | 0.8339 | 0.7201 |
| | ViT | 0.7510 | 0.8277 | 0.7165 |
| | Swin | 0.8099 | 0.8630 | 0.8144 |
| XLM-R | CLIP | 0.8097 | 0.8664 | 0.7917 |
| | BEiT | 0.8016 | 0.8634 | 0.7661 |
| | ViT | 0.6728 | 0.7557 | 0.6848 |
| | Swin | 0.8014 | 0.8658 | 0.7494 |
| **BanglaBERT-2** | **CLIP** | **0.8283** | **0.8802** | **0.8066** |
| | BEiT | 0.8189 | 0.8741 | 0.7930 |
| | ViT | 0.8048 | 0.8655 | 0.7709 |
| | Swin | 0.7982 | 0.8588 | 0.7771 |

Table 4: Performance of multimodal combinations.

Following thorough hyperparameter tuning, the proposed model reached an M-F1 score of 85.36%. This represents a 5.02% improvement over the best text model (BanglaBERT-1, 80.34%), a 6.42% gain over the best visual model (CLIP, 78.94%), and a 7.02% increase compared to the best multimodal baseline (CLIP, 78.34%).

### 5.1.3 Impact of pre-trained multimodal baselines

Several prebuilt multimodal models were evaluated, including BLIP-2, CLIP, M-CLIP, and VisualBERT. Table 5 presents their performance metrics. However, these prebuilt solutions consistently underperformed relative to the custom fusion-based multimodal systems. Notably, CLIP achieved the highest M-F1 score among them at 78.34%, which remains significantly lower than that of the proposed method (BanglaBERT-2 + CLIP). This marked difference highlights the clear superiority of carefully designed fusion strategies over generic, end-to-end pre-trained multimodal models for classifying propagandistic memes in Bengali.

| Model | M-F1 | W-F1 | G |
|---|---|---|---|
| BLIP-2 | 0.7300 | 0.8200 | 0.6425 |
| **CLIP** | **0.7834** | 0.8491 | 0.7577 |
| M-CLIP | 0.7430 | 0.8032 | 0.7838 |
| VisualBERT | 0.5755 | 0.6636 | 0.6034 |

Table 5: Performance of pre-trained multimodal models.

### 5.1.4 Impact of hyperparameters' tuning on performance

All models were trained using 70% of the dataset, validated on 15%, and tested on the remaining 15%. We conducted extensive hyperparameter tuning on the top-performing model (e.g., BanglaBERT-2+CLIP) to further enhance its performance. This process aimed to identify the optimal configuration to improve classification results and robustness. Table 6 provides a detailed overview of the selected hyperparameters and their impact on performance metrics. After tuning, with data split (80-10-10), late fusion, learning rate (5e-5), batch size (4), weight decay (0.1), and training for 20 epochs with gamma=2, the BanglaBERT-2+CLIP model achieved an M-F1 of 85.36%, about 2.53% higher than the initial configuration (82.83%), with corresponding improvements in W-F1 (89.86%) and G-Mean (82.97%), demonstrating that careful hyperparameter optimization significantly boosts performance.

| Hyperparameter | Optimal | M-F1 | W-F1 | G |
|---|---|---|---|---|
| Data Split {60, 70, 80} | 80-10-10 | 0.8324 | 0.8777 | 0.8523 |
| Fusion Type {Feature, Late} | Late | 0.8356 | 0.8843 | 0.8257 |
| LR {(1,2,5) e-5, 5e-4} | 5e-5 | 0.8406 | 0.8887 | 0.8236 |
| Batch Size {4, 8, 16} | 4 | 0.8536 | 0.8986 | 0.8297 |
| WD {0.01, 0.1} | 0.1 | 0.8536 | 0.8986 | 0.8297 |
| Gamma & Epochs | 2, 20 | **0.8536** | **0.8986** | **0.8297** |

Table 6: Performance across different hyperparameter configurations for BanglaBERT-2+CLIP model.

### 5.1.5 Comparison with existing techniques

To evaluate the effectiveness of the proposed model, we benchmarked its performance against several existing multimodal approaches (Zaytoon et al., 2024; Hasanain et al., 2024b; Qu et al., 2022) on the dataset we developed. Table 7 shows that the proposed method achieved the highest W-F1 score of 89.86%, surpassing all existing techniques and demonstrating superior capability in propagandistic meme detection across modalities. The proposed method outperforms the second-best approach (Bloomz-1b1+ ResNe101) by achieving an absolute improvement of 3.05% in M-F1 (from 85.36% to 88.41%) and 2.55% in W-F1 (from 89.86% to 92.41%).

| Model | M -F1 | W-F1 | G |
|---|---|---|---|
| Bloomz-1b1 + ResNet101 (Zaytoon et al., 2024) | 0.8231 | 0.8731 | **0.8316** |
| ResNet + BERT to SVM (Hasanain et al., 2024b) | 0.7293 | 0.8270 | 0.6243 |
| VisualBERT-COCO (Qu et al., 2022) | 0.6233 | 0.7016 | 0.6648 |
| CLIP (Li et al., 2024) | 0.7903 | 0.8509 | 0.7854 |
| **Proposed** (BanglaBERT-2+CLIP) | **0.8536** | **0.8986** | 0.8297 |

Table 7: Benchmarking of multimodal models on the test set.

### 5.1.6 Ablation Study

To assess the contributions of image-text modalities and fusion techniques, we perform an ablation study utilizing the macro F1 score. Text and image modalities perform competitively independently, with BanglaBERT-1 slightly outperforming CLIP (80.34% vs. 78.94%). Combining both using feature-level fusion—where representations from each modality are merged before the final classification—increases performance to 82.83%, demonstrating the complementary nature of the two approaches. The proposed late-fusion method integrates BanglaBERT-2 and CLIP by maintaining separate modality-specific representations and combining them only at the decision level, with improved hyperparameter tuning, which achieves the best macro F1 score of 85.36%. This represents gains of +2.53 points over intermediate fusion and +5.02 / +6.42 points compared to the text-only and image-only baselines, respectively, demonstrating that preserving modality-specific representations and integrating them later yields superior results for multimodal propagandistic meme detection.

## 5.2 Error Analysis

To gain an in-depth understanding of the proposed model's performance, a thorough error analysis is conducted using both quantitative and qualitative methods. The following parts present a detailed error analysis of the BanglaBERT-2+CLIP model.

### 5.2.1 Quantitative Error Analysis

The confusion matrix (Figure 4) confirms strong classification performance (337/374 correct; 90.11% accuracy). For the positive class, the error rate is 27.9% with 24 misclassified samples; for the negative class, the error rate is 4.5% with 13 misclassified samples, while the per-class weighted F1-scores are 0.7214 for *Non-prop* and 0.1771 for *Prop*, which sum to the overall weighted average of 0.8986. The lower weighted F1-score for propagandistic (0.1771) compared to non-propagandistic (0.7214) is primarily due to the severe class imbalance in the dataset, where class 1 accounts for only 23.1% (86 samples) of the total data. In comparison, class 0 dominates with 76.9% (288 samples). There are low false positives (4.5%) but higher false negatives (27.9%), indicating high precision yet room for improvement in recall. Overall, the results highlight the potential of multimodal transformer models for propagandistic content detection in multilingual contexts.

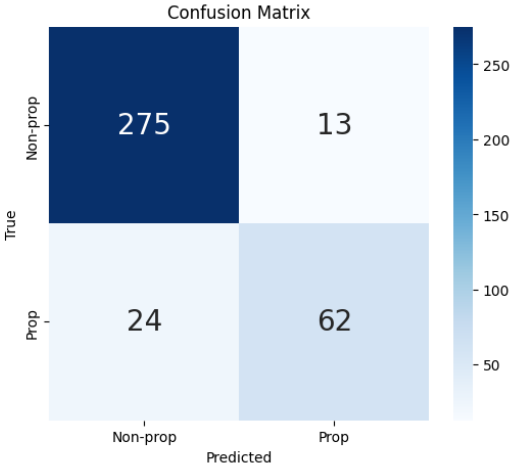
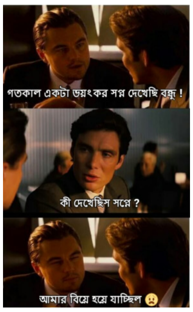


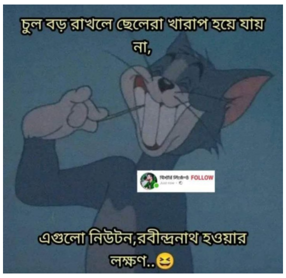
Figure 4: Confusion matrix of the proposed model.

## 5.3 Qualitative Error Analysis

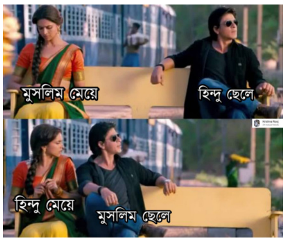
Figure 5 shows representative examples of correct and incorrect predictions generated by the proposed method (BanglaBERT-2+CLIP) compared with BanglaBERT-2 and CLIP. In Figure 5a, all three models correctly labelled the meme as non-propagandistic, indicating strong agreement between visual and textual modalities when both cues align. In Figure 5b, the textual model misclassified the humorous content as propagandistic, and the multimodal model reflected this textual bias, whereas the visual model relied on image cues and correctly identified it as non-propagandistic, highlighting each model's dependency on its respective modality. Figure 5c shows a challenging case where the textual model detected propagandistic content, but the visual and multimodal models did not, due to reliance on visual cues that missed communal or cultural propaganda. The ground truth confirmed the propagandistic label, underscoring that models that prioritize visual context may overlook textual signals. Finally, Figure 5d shows that all modalities converged on the correct propagandistic classification, demonstrating that explicit textual propaganda, when supported visually, enables consistent predictions based on combined cues.



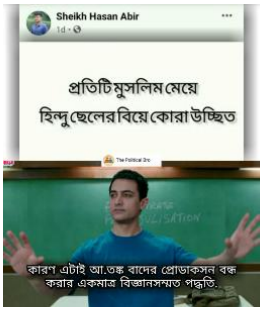
(a) Actual: Non-prop √
Textual: Non-prop √
Visual: Non-prop √
Multimodal: Non-prop √


চুল বড় রাখলে ছেলেরা খারাপ হয়ে যায় না,
FOLLOW
এগুলো নিউটন,রবীন্দ্রনাথ হওয়ার লক্ষণ..

(b) Actual: Non-prop
Textual: Prop X
Visual: Non-prop √
Multimodal: Prop X



(c) Actual: Prop
Textual: Prop √
Visual: Non-prop X
Multimodal: Non-prop X


Sheikh Hasan Abir
1d
প্রতিটি মুসলিম মেয়ে
হিন্দু ছেলের বিয়ে কোরা উচ্ছিত
The Political Bro
কারণ এটাই আ.তঙ্ক বাদের প্রোডাকসন বন্ধ করার একমাত্র বিজ্ঞানসম্মত পদ্ধতি.

(d) Actual: Prop √
Textual: Prop √
Visual: Prop √
Multimodal: Prop √

Figure 5: Examples of correct and incorrect predictions by the proposed model.

The four meme samples illustrate how each model's reliance on either textual, visual, or multimodal cues affects prediction accuracy. For instance, when both text and image suggest harmless humour, predictions are generally reliable across all models. However, challenges arise when propagandistic messages are covertly embedded in images that textual models may miss, or when culturally sensitive language in text prompts false positives from text-dependent models. These modality dependencies complicate the distinction between propaganda and non-propaganda, especially when nuanced religious or political content is present. Such subtleties can confuse models, leading to errors when one modality dominates interpretation. This dependency limits the models' ability to generalize across diverse meme formats, often leading to misclassification in emotionally or culturally complex instances and underscoring the importance of integrating and balancing multiple modalities for comprehensive understanding.

## 6 Conclusion

This work presents **MemeGuard**, a new dataset for detecting propaganda in Bengali memes. Using this dataset, forty-five unimodal and multimodal models are systematically evaluated for this task. Evaluation demonstrates that the BanglaBERT-2+CLIP model decisively surpasses all unimodal and multimodal baselines after fine-tuning on MemeGuard, achieving the top macro F1 score (85.36%) and weighted F1 (89.86%). These results highlight the strength of the proposed multimodal fusion in identifying propagandistic content. Future research will address current model limitations by expanding and diversifying the dataset, enhancing code-mixed data handling, exploring cutting-edge multimodal architectures such as LLMs and VLMs, refining fine-grained propaganda detection, and implementing automated hyperparameter optimization, while also leveraging a mix of our limited labeled corpus (3,745 samples) and additional unlabeled memes to enable semi-supervised learning, allowing the model to exploit abundant unlabeled data for richer representation learning—an approach particularly valuable for low-resource languages like Bengali. Beyond the well-established late-fusion strategy that yielded our best performance, future directions include investigating alternative fusion mechanisms, such as cross-modal attention, early or hierarchical fusion, and adaptive or gated fusion, as well as modality-specific architectural innovations, such as improved vision encoders, code-mixed language models, and joint embedding frameworks for more substantial multimodal alignment.

## Limitations

Although the proposed method performs well at detecting propaganda in memes, several significant limitations remain unaddressed.

- A limited dataset size and narrow data sources may reduce generalizability.
- Class imbalance, with only 23.1% of samples labelled as propagandistic, may introduce bias in model training.
- Manual hyperparameter tuning is time-intensive and may not produce optimal results.
- Focusing solely on Bengali memes limits the method's applicability to other languages.

## Ethics Statement

This study uses memes exclusively for academic research, without the intent to promote or distribute harmful material. Dataset creation and annotation adhered to copyright regulations, protected annotator privacy, and ensured fairness. Multiple annotators participated to minimize bias; however, cultural subjectivity persists as a limitation. The proposed system serves as a research instrument rather than a censorship mechanism, and human oversight remains necessary for practical implementation.

## Declaration of AI Tools Uses

This manuscript complies with ACL policies on ethical AI use. ChatGPT (OpenAI) and Grammarly were used only for language refinement, such as improving grammar, clarity, coherence, and writing quality. The tools were not used to generate scientific ideas, analyze data, interpret results, or draw conclusions.

## Acknowledgment

This work was supported by the Directorate of Research & Extension (DRE), Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh, under the grant number: CUET/DRE/2023-2024/CSE/024.

## References

2025. Multibanfakedetect: Integrating advanced fusion techniques for multimodal detection of bangla fake news in under-resourced contexts. *International Journal of Information Management Data Insights*, 5(2):100347.

Tanzin Ahammad, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2025. M-sam: Multimodal sentiment analysis exploiting textual and visual features of social media memes. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 134–150, Cham. Springer Nature Switzerland.

Firoj Alam, Abul Hasnat, Fatema Ahmed, Md Arid Hasan, and Maram Hasanain. 2024a. Armeme: Propagandistic content in arabic memes. *arXiv preprint arXiv:2406.03916*.

Md Ashraful Alam, Jawad Hossain, Shawly Ahsan, and Mohammed Moshiul Hoque. 2024b. Multimodal aggressive meme classification using bidirectional encoder representations from transformers. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3542–3547. IEEE.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Ziyang Cheng. 2025. Internet meme culture and political propaganda: The impact of 2024 us election memes on chinese online community. *Media Watch*, page 09760911251368965.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Avishek Das, Moumita Sen Sarma, Mohammed Moshiul Hoque, Nazmul Siddique, and M. Ali Akber Dewan. 2024. Avater: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition. *Sensors*, 24(18).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Md Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero-and few-shot prompting with llms: A comparative study with fine-tuned models for bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818.

Md. Maruf Hasan, Shawly Ahsan, Mohammed Moshiul Hoque, and M. Ali Akber Dewan. 2025. Mulad: Multimodal aggression detection from social media memes exploiting visual and textual features. In *Pattern Recognition*, pages 107–123, Cham. Springer Nature Switzerland.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Md Arid Hasan, Fatema Ahmed, Reem Suwaileh, Md Rafiul Biswas, Wajdi Zaghouani, and Firoj Alam. 2024b. Araieval shared task: propagandistic techniques detection in unimodal and multimodal arabic content. *arXiv preprint arXiv:2407.04247*.

Aleš Horák, Radoslav Sabol, Ondřej Herman, and Vít Baisa. 2024. Recognition of propaganda techniques in newspaper texts: Fusion of content and style analysis. *Expert Systems with Applications*, 251:124085.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing: student research workshop*, pages 32–39.

Md Rajib Hossain, Sadia Afroze, Asif Ekbal, Mohammed Moshiul Hoque, and Nazmul Siddique. 2025. Multimodfusenet: Advancing multimodal text classification for low-resource languages through textual-visual feature fusion. *Knowledge-Based Systems*, 328:114085.

Chinmaya Hs and 1 others. 2021. Trollmeta@ dravidianlangtech-eacl2021: Meme classification using deep learning. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 277–280.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul NC, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. Memetector: Enforcing deep focus for meme detection. *International Journal of Multimedia Information Retrieval*, 12(1):11.

Shiyi Li, Yike Wang, Liang Yang, Shaowu Zhang, and Hongfei Lin. 2024. Lmeme at semeval-2024 task 4: Teacher student fusion-integrating clip with llms for enhanced persuasion detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 628–633.

Mohamed Lichouri, Khaled Lounnas, Aicha Zitouni, Houda Latrache, and Rachida Djeradi. 2023. Usthb at araieval'23 shared task: Disinformation detection system based on linguistic feature concatenation. In *Proceedings of ArabicNLP 2023*, pages 508–512.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Tarek Mahmoud and Preslav Nakov. 2024. Bertastic at semeval-2024 task 4: State-of-the-art multilingual propaganda detection in memes via zero-shot learning with vision-language models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 503–510.

Monsoon NLP. 2024. Bangla-electra: A pretrained electra model for bengali. `https://huggingface.co/monsoon-nlp/bangla-electra`. Accessed: 2025-06-25.

Pir Ahmad Noman, Liu Yuanchao, Khursheed Aurangzeb, Muhammad Anwar Shahid, and Qazi Mazhar ul Haq. 2024. Semantic web-based propaganda text detection from social media using meta-learning. *Service Oriented Computing and Applications*.

Olumide E Ojo, Olaronke O Adebanji, Hiram Calvo, Damian O Dieke, Olumuyiwa E Ojo, Seye E Akinsanya, Tolulope O Abiola, and Anna Feldman. 2023. Legend at araieval shared task: Persuasion technique detection using a language-agnostic text representation model. *arXiv preprint arXiv:2310.09661*.

C.A. Piña-García. 2025. In-context learning for propaganda detection on twitter mexico using large language model meta ai. *Telematics and Informatics Reports*, 19:100232.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *arXiv preprint arXiv:2205.12617*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Md. Tanvir Rahman, Shawly Ahsan, Jawad Hossain, Mohammed Moshiul Hoque, and M. Ali Akber Dewan. 2025. Multimodal emotion recognition system leveraging decision fusion with acoustic and visual cues. In *Pattern Recognition. ICPR 2024 International Workshops and Challenges*, pages 117–133, Cham. Springer Nature Switzerland.

Muhammad Umar Salman, Asif Hanif, Shady Shehata, and Preslav Nakov. 2023. Detecting propaganda techniques in code-switched social media text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16794–16812, Singapore. Association for Computational Linguistics.

Sagor Sarker. 2021. Banglabert: A bert-based language model for bengali. `https://huggingface.co/sagorsarker/bangla-bert-base`. Accessed: 2025-06-25.

Anik Mahmud Shanto, Mst. Sanjida Jamal Priya, Fahim Shakil Tamim, and Mohammed Moshiul Hoque. 2025. MDC$^3$: A novel multimodal dataset for commercial content classification in Bengali. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 311–320, Albuquerque, USA. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Ming-Hung Wang and Yu-Lin Chen. 2025. Beyond text: Detecting image propaganda on online social networks. *IEEE Transactions on Sustainable Computing*, 10(1):120–131.

Mohamed Zaytoon, Nagwa M El-Makky, and Marwan Torki. 2024. Alexunlp-mz at araieval shared task: contrastive learning, llm features extraction and multi-objective optimization for arabic multi-modal meme propaganda detection. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 512–517.

Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal understanding of memes with fair explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2007–2017.

## A Annotation Guidelines

The following questions, tied to specific propaganda characteristics, guide the classification process, where memes that fall under one or more of these characteristics are considered propagandistic, and those that do not are treated as non-propagandistic:

- **Intent to Influence or Manipulate**: (i) Does the meme push a specific political, ideological, or social agenda? (ii) Does it encourage the viewer to act (support a cause, oppose a group, or adopt a belief)?
- **Emotional Appeal**: (i) Does the meme evoke strong emotions like fear, anger, pride, or sympathy? (ii) Does it use fear, exaggeration, or threats to influence opinions or actions?
- **Simplification of Complex Issues**: Does the meme reduce a complex issue to overly simple terms?
- **Polarization and Division**: Does the meme create an "us versus them" narrative?
- **Repetition and Catchphrases**: Does the meme repeat messages or use catchy slogans that stick in the audience's mind?
- **Misleading Information**: Does the meme include misinformation or disinformation?
- **Smear Tactics**: Does the content use negative claims to undermine the reputation of an individual or group without providing credible evidence?
- **Visual Symbolism and Transfer**: Does the meme use images or symbols (like national flags, religious icons, or culturally significant visuals) to evoke specific associations or emotions?

## B Cohen's Kappa Score

The annotation quality for the MemeGuard dataset was assessed using Cohen's kappa coefficient among three undergraduate annotators, as shown in Table 8. Pairwise kappa scores were 0.85 (Annotator 1 & 2), 0.90 (Annotator 1 & 3), and 0.77 (Annotator 2 & 3), indicating significant to almost perfect agreement. The average kappa score of 0.84 shows the high consistency and reliability of the annotations, highlighting the robustness of the dataset for detecting Bengali propagandistic memes.

| **Pair** | **Kappa Score** |
|---|---|
| P-1 | 0.85 |
| P-2 | 0.90 |
| P-3 | 0.77 |
| Average | 0.84 |

Table 8: Pairwise Cohen's kappa score

## C Hyperparameters

Table 9 provides a detailed overview of the hyperparameters chosen for training both unimodal and multimodal baselines. All models were trained using these values, which we examined across every model.

| **Hyperparameter** | **Search Space** |
|---|---|
| Batch Size | 4, 8, 16 |
| Epochs | 10, 15, 20 |
| Optimizer | Adam |
| Weight Decay | 0.01, 0.1 |
| Learning Rate | 5e-4, (1, 2, 5) e-5 |

Table 9: Hyperparameters for all models

# Language as a Label: Zero-Shot Multimodal Classification of Everyday Postures under Data Scarcity

**MingZe Tang, Jubal Chandy Jacob**
Department of Computing Science
University of Aberdeen
{m.tang.24, j.chandyjacob.24}@abdn.ac.uk

## Abstract

This paper investigates how the specificity of natural language prompts influences zero-shot classification performance in modern vision language models (VLMs) under severe data scarcity. Using a curated 285 image subset of MS COCO containing three everyday postures (sitting, standing, and walking/running), we evaluate OpenCLIP, MetaCLIP 2, and SigLIP alongside unimodal and pose-based baselines. We introduce a three tier prompt design, minimal labels, action cues, and compact geometric descriptions and systematically vary only the linguistic detail. Our results reveal a counterintuitive trend where simpler prompts consistently outperform more detailed ones, a phenomenon we term *prompt overfitting*. Grad-CAM attribution further shows that prompt specificity shifts attention between contextual and pose-relevant regions, explaining the model dependent behaviour. The study provides a controlled analysis of prompt granularity in low resource image based posture recognition, highlights the need for careful prompt design when labels are scarce.

## 1 Introduction

Label scarcity is a central barrier for practical human action recognition from still images (Wu et al., 2022). Many deployments cannot acquire balanced annotations or run task specific training. Vision and language encoders mitigate this limitation by learning a shared embedding space in which text can serve as a label at inference time (Radford et al., 2021). This paper studies whether careful wording of those text labels improves zero shot classification under data scarcity.

The task focuses on three everyday postures in still images, namely sitting, standing, and walking or running, using a small subset derived from COCO (Lin et al., 2014) with 230 images. Image content, preprocessing, and scoring are held fixed, and language acts as the only supervision at inference. Each image is embedded once at the native input size of the model and is scored by cosine similarity against one prompt per class. Prompt specificity is the sole experimental factor and follows a three tier design. Tier one uses a minimal label template. Tier two adds a short action cue. Tier three adds compact pose geometry that specifies body configuration. Prompts exclude scene, identity, and clothing terms so that differences arise only from pose description.

Evaluation covers multimodal encoders that align images and text, namely OpenCLIP, MetaCLIP and SigLIP. Vision only baselines include DINOv3 and a standard Vision Transformer paired with frozen sentence embeddings to form a heuristic zero shot classifier. A pose based baseline uses YOLOv11 Pose for key-point estimation together with a simple geometric decision rule. Results are reported as accuracy and macro F1 for each tier and each model, and qualitative analysis with gradient based visualisations assesses whether greater prompt specificity shifts attention toward pose relevant regions. The study provides an empirical protocol for zero shot recognition under data scarcity and a controlled comparison of prompt wording across modern encoders and non linguistic baselines.

## 2 Related Works

### 2.1 Vision–Language Models for Zero-Shot Classification

At the core of modern zero-shot classification is the contrastive language–image pre-training paradigm introduced by CLIP, which aligns visual and textual representations in a shared embedding space using large collections of image–text pairs (Radford et al., 2021; Jia et al., 2021). The objective draws paired images and texts closer while separating mismatched pairs, thereby encoding vision–language

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 48–57
December 23, 2025 ©2025 Association for Computational Linguistics

correspondences (Zhang et al., 2023). Zero-shot classification then becomes a nearest-neighbour search in this shared space: given an input image and a set of class descriptions, the model computes cosine similarity between image features and the text embeddings of prompts such as “a photo of a `[class]`”, predicting the class with the highest similarity (Ghiasvand et al., 2025b,a).

OpenCLIP demonstrates competitive results across more than thirty benchmarks spanning OCR, scene recognition, and fine-grained object categorisation, often approaching supervised baselines without task-specific training (Radford et al., 2021). The contrastive formulation has since been adapted for downstream tasks including detection, segmentation, video action recognition, and depth estimation (Xu et al., 2023; Zhou et al., 2022; Xu et al., 2022). More recent models such as SigLIP revisit the pre-training loss, replacing the softmax contrastive objective with independent sigmoid scoring (Zhai et al., 2023), while MetaCLIP 2 improves performance by scaling and curating training data (Chuang et al., 2025). Dual-encoder VLMs remain dominant due to their scalability, robustness, and task flexibility (Volkov et al., 2025; Zhang et al., 2023).

### 2.2 Low-Resource & Low-Compute Image Understanding

Balancing performance and computational efficiency has motivated training-free or parameter-efficient approaches for resource-constrained environments (Zhang et al., 2024). These methods leverage pre-trained representations to extract more information from test samples and class names without parameter updates a valuable property when labelled data, compute, or both are limited. Parameter-efficient adaptation, including prompt tuning and related techniques, offers a middle ground by updating only a small subset of parameters or learned tokens while keeping backbone weights frozen (Mistretta et al., 2024; Lester et al., 2021). Training-free few-shot methods such as TIP-Adapter further exploit cached support embeddings to merge zero-shot textual knowledge with visual evidence (Esbri et al., 2024).

Resource constraints also arise from domain shift and model scale. Direct zero-shot deployment of large VLMs can be hindered by distributional mismatch, particularly in specialised settings such as medical imaging (Wang et al., 2025; Liu et al., 2023). Video understanding presents additional challenges, as spatiotemporal modeling increases computation demands (Bosetti et al., 2024; Shao et al., 2020). Collectively, these findings highlight the importance of methods that maximise generalisation under minimal supervision. VLMs function effectively as “data multipliers,” transferring broad, open-world knowledge into low-resource classification settings (Zhang et al., 2024; Volkov et al., 2025). Our study takes this perspective by comparing multimodal and vision-only models under severe data scarcity.

### 2.3 Prompt Semantics, Granularity, and Posture Cues

Early work on domain-specific zero-shot recognition used manually specified semantic attributes to represent actions or activities, enabling recognition of unseen classes through structured linguistic or conceptual descriptions (Bosetti et al., 2024; Zellers and Choi, 2017). Later approaches replaced hand-built attributes with distributional word embeddings, framing zero-shot recognition as alignment between visual features and latent semantic spaces. These ideas extend naturally to video settings, where language acts as the primary mechanism for generalisation (Bosetti et al., 2024; Shao et al., 2020).

Prompt engineering has since become a central research focus, especially as models exhibit varying sensitivity to how class concepts are phrased. Recent studies show that prompt wording measurably affects zero-shot action recognition; prompts generated by large language models frequently outperform raw label prompts (Ali et al., 2024). Methods such as CuPL automate descriptive prompt construction and demonstrate consistent improvements on benchmarks like ImageNet (Cai et al., 2025; Pratt et al., 2022). However, Xu et al. highlight that increasing linguistic specificity does not universally improve model performance and may instead introduce distributional mismatches between text and visual cues (Xu et al., 2023).

This motivates the need to understand how multimodal encoders respond to differing levels of semantic granularity, especially for fine-grained distinctions such as human postures. Our study builds directly on this line of research: rather than generating long-form descriptions, we systematically manipulate prompt specificity minimal labels, action cues, and compact geometric wording to examine how VLMs ground pose semantics under extreme data scarcity. This controlled setup allows us to identify a counterintuitive phenomenon we term *prompt overfitting*, where excessive detail harms performance for stronger models.

## 3 Data and Methods

### 3.1 Dataset

We evaluate on a curated 285–image subset of MS COCO Lin et al. (2014), a large–scale benchmark of everyday scenes with dense instance annotations. From the 2014 releases, images containing at least one visible person with sufficient visual evidence to judge posture were sampled, and a single action label was assigned per image by manual inspection. The subset is balanced across three classes: *sitting* ($n$=95), *standing* ($n$=92), and *walking or running* ($n$=98). Figure 1 shows randomly selected examples for each class and illustrates variation in viewpoint, background, and occlusion. Exploratory analysis of raw image sizes showed a concentration around 640×480 pixels, which supports a uniform resize to 224×224 for all models. Aspect–ratio distributions did not differ across classes, and visual checks confirmed negligible class–specific distortion after resizing.

Figure 1: Random samples from the curated MS COCO subset for *sitting*, *standing*, and *walking or running*.

### 3.2 Models and Experimental Approach

Our evaluation is structured around three distinct representation paradigms. In each case, the pre-trained model serves as a feature extractor, with a lightweight classifier trained on the resulting embeddings.

#### 3.2.1 Unimodal Vision Models

**Vision Transformer (ViT)** A pre-trained ViT (Sreekanth, 2024) model (`vit-base-patch16-224`) was fine-tuned on the dataset. It was evaluated on both a binary task (sitting vs. standing) and the full three-class task to assess its generalization capability.

**DinoV3** A state-of-the-art vision model pre-trained using self-supervised learning on images alone, allowing us to assess the efficacy of purely visual, non-linguistic representation learning (Siméoni et al., 2025).

#### 3.2.2 Multimodal Vision-Language Models

**OpenCLIP** We employed OpenAI (Radford et al., 2021)'s CLIP pre-trained vision encoder `openai/clip-vit-base-patch32` model, leveraging its extensive image-text pre-training to generate semantically rich embeddings.

**Meta CLIP 2** An evolution of CLIP, pre-trained on a more meticulously curated dataset to enhance the quality and robustness of its visual-semantic representations (**?**).

**SigLip** A VLM employing a sigmoid-based loss function during pre-training, offering an alternative to the contrastive objective of CLIP (**?**).

#### 3.2.3 Pose-Centric Structural Model

**YOLOv11x-pose** This model implements a two-stage process. First, the YOLOv11x pose (Khanam and Hussain, 2024) architecture is applied to each image to extract a set of 2D keypoints that represents the subject's skeleton. Second, geometric features, such as the angles between the left and right knee and hip joints, are calculated from these key points. A simple classifier is then trained on these angular features to determine the final action class.

### 3.3 Prompt tiering for zero-shot classification

We vary only the specificity of the text prompt in order to test how wording affects zero-shot posture recognition with scarce data. **Tier 1** uses the class label in a minimal template such as "a photo of a person `[class]`", which reflects common zero-shot practice. **Tier 2** adds a brief action cue that clarifies the target category, for example "a person seated on a chair", "a person standing still and upright", or "a person mid-stride with one foot off the ground". **Tier 3** replaces action words with short anatomical or pose constraints, for example "hips and knees bent at right angles" for sitting or "legs straight and torso vertical" for standing. Across tiers we keep prompts scene free and we avoid background, clothing, and identity terms so that only pose information differs. For each tier we create one prompt per class, compute unit-normalized text embeddings once, embed each image once at the model's native resolution, and score classes by cosine similarity. We report accuracy and macro-F1 per tier on the same images and preprocessing settings without any model fine-tuning so that observed differences can be attributed to prompt content rather than changes in data or optimization.

### 3.4 Experimental Approach

All experiments were conducted with a focus on reproducibility and were run on Google Colab on a single NVIDIA T4 GPU with 16GB of memory. The dataset was partitioned on a fixed stratified 80% training, 10% validation, and 10% test split for all experiments. A global random seed was established to ensure that all models were trained and evaluated on the exact same data partitions.

The task was defined in two distinct classification scenarios to assess the performance of the model at varying levels of difficulty: (1) A simplified binary task focusing on two more visually distinct classes: *sitting* vs. *walking/running*. (2) Three-class task encompassing all labels: *sitting*, *standing*, and *walking/running*; To account for stochasticity in the training process, each model was trained and evaluated over five independent runs with different seeds. We used an early stopping mechanism with patience of 5 epochs, monitoring the validation loss to prevent over-fitting.

## 4 Results

The empirical evaluation is presented in two parts. (1) A comparative analysis of different model architectures under a standard training and evaluation paradigm to establish baseline performance. (2) A zero-shot experiment investigating how the specificity of text prompts affects the performance of Vision-Language Models.

### 4.1 Comparative Analysis of Model Architectures

A comparative evaluation of models from three different paradigms (Unimodal, VLM, and Pose-Centric) was conducted across both binary and multi-class classification tasks as the first stage of the evaluation.

#### 4.1.1 Performance on Binary Classification

To investigate how model performance is affected by task complexity, the models were evaluated on a binary classification task (sitting vs. walking/running). The results, presented in Table 1, show a general performance uplift across most models, yet the relative ranking remains largely consistent.

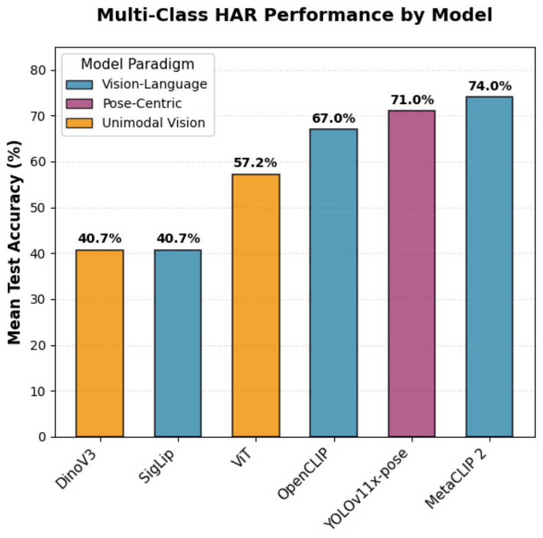


Figure 2: Mean test accuracy on the primary three-class HAR task. Models with semantic (VLM) or structural (Pose) priors demonstrate a clear performance advantage.

MetaCLIP 2 performed the best with an accuracy of 92.8%. Notably, ViT performed well in this less ambiguous setting, achieving 90.0% accuracy, nearly matching the top VLM. This suggests that when classes are more visually distinct, a powerful unimodal architecture can be highly effective. The original OpenCLIP model also performed strongly at 88.1%. However, DinoV3 and SigLip surprisingly continued to lag significantly, with accuracies of 57.5% and 56.5%, respectively.

#### 4.1.2 Performance on Multi-Class Classification

The three-class action recognition task represents the core challenge of this study, requiring models to distinguish between visually similar and ambiguous postures from a single static frame. As detailed in Figure 2, the performance of the evaluated models diverged significantly, clearly separating them into distinct tiers.

The models endowed with strong priors formed the top tier. The Vision-Language Model MetaCLIP 2 achieved the highest accuracy at 74.0%. Following closely was the YOLOv11x-pose model, which, by leveraging a structural representation of the human body, secured an accuracy of 71.0%. The original OpenCLIP model also delivered a robust performance of 67.0%.

A substantial performance gap exists between these models and the unimodal models that learn from pixels alone. The standard Vision Transformer (ViT) achieved a modest accuracy of 57.2%. The purely self-supervised DinoV3 and the VLM SigLip both struggled significantly, each attaining only 40.7% accuracy, a result only marginally better than random chance.

| Model | B. Acc. | M. Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|
| MetaCLIP 2 | 0.92 | 0.74 | 0.74 | 0.74 | 0.74 |
| ViT | 0.90 | 0.52 | 0.59 | 0.57 | 0.57 |
| OpenCLIP | 0.88 | 0.67 | 0.68 | 0.67 | 0.66 |
| YOLOv11x-pose | — | 0.71 | 0.73 | 0.71 | 0.71 |
| DinoV3 | 0.57 | 0.40 | 0.41 | 0.41 | 0.40 |
| SigLip | 0.56 | 0.40 | 0.28 | 0.41 | 0.33 |

Table 1: Performance comparison of vision models, reporting **(B. Acc.)** Binary Accuracy, **(M. Acc.)** Multi-class Accuracy, **(Prec.)** Macro Precision, **(Rec.)** Macro Recall, and **(F1)** Macro F1 Score.

#### 4.1.3 Analysis of Class-Specific Metrics

To gain a more nuanced understanding, we analyzed the macro-averaged Precision, Recall, and F1-Score (Table 1). These metrics reinforce the hierarchy observed in accuracy. MetaCLIP 2 and YOLOv11x-pose demonstrated a strong balance between precision and recall, resulting in high F1-Scores of 0.74 and 0.71, respectively, indicating reliable classification across all three categories. In contrast, lower-performing models exhibited imbalances. For instance, SigLip2 had a recall of 0.41 but a very low precision of 0.28, suggesting it generated a large number of false positive predictions in its attempt to classify instances from all classes.

### 4.2 Prompt-Specific Zero-Shot Performance

In our second set of experiments, we investigated how prompt specificity affects the zero-shot performance of VLMs. The results, detailed in Table 2, show that the relationship between prompt detail and model performance is not linear and is highly model-dependent.

#### 4.2.1 Performance Trends for High Performing Models

The primary trend observed for the leading Vision-Language Models is a clear inverse relationship between prompt specificity and classification performance. As represented in Table 2 both MetaCLIP 2 and OpenCLIP, the simplest Tier 1 prompts consistently achieved the highest accuracy and F1 scores. The introduction of more descriptive features in Tier 2 or anatomical cues in Tier 3 resulted in a significant degradation of performance. This effect was particularly pronounced for MetaCLIP 2, where the multi-class accuracy fell sharply from 68.8% with a Tier 1 prompt to 55.1% with a Tier 2 prompt. Similarly, OpenCLIP's multi-class accuracy saw a substantial decrease from a high of 71.2% (Tier 1) to 52.6% (Tier 2). This consistent impact suggests a phenomenon of "prompt overfitting" where excessive detail may unduly constrain the models and hinder their ability to generalize.

#### 4.2.2 Model-Dependent Responses to Prompt Granularity

The Tier 1 performance trend was not universal, highlighting that the optimal prompt strategy is highly model dependent. The lower-performing SigLip model exhibited a contrasting response to the increase in prompt detail. While its overall accuracy remained consistently low, its ability to classify the ambiguous `walking_running` class was significantly boosted by the specific, "body cue-based" Tier 3 prompts. This was most evident in the binary task, where the F1 score for this specific class jumped from 0.364 with a basic Tier 1 prompt to 0.566 with the detailed Tier 3 prompt.

## 5 Discussion & Conclusion

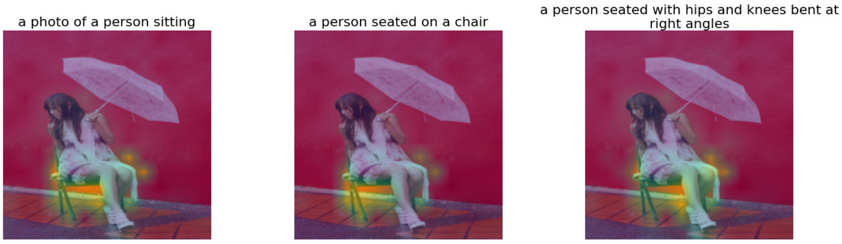
### 5.1 Prompt Specificity as Supervision at Inference

Prompt wording functions as an explicit prior on the classifier decision in a zero-shot setting. Minimal, noun-centric prompts align with the distributions seen during pre-training of image–text encoders, where concept names are frequent and broadly grounded. This alignment explains the competitiveness of label-only prompts in closed sets. In contrast, adding brief action cues can introduce a linguistic–visual mismatch for still images, since verbs such as "walking" or "standing still" denote dynamics or intent rather than stable appearance. The resulting text embeddings are drawn toward contexts that are weakly supported by a single frame, which reduces similarity margins and increases overlap between neighbouring classes.

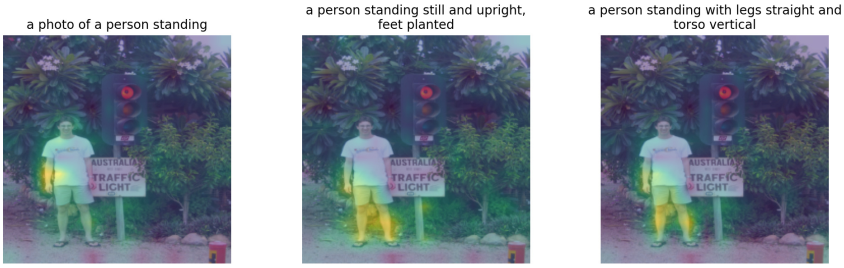
Geometric phrasing exerts a different influence. Short anatomical constraints specify local, view-stable relations, for example relative angles at the hip and knee or verticality of the torso, that are directly verifiable in a single image. Gradient-based attributions (see Figure 3a) consistently show increased concentration over limb and torso regions

| **Task** | **Tier** | **MetaCLIP 2** | | **OpenCLIP** | | **SigLIP** | |
|---|---|---|---|---|---|---|---|
| | | **Accuracy** | **Macro F1** | **Accuracy** | **Macro F1** | **Accuracy** | **Macro F1** |
| **Binary** | Tier 1 | 0.938 | 0.938 | 0.907 | 0.907 | 0.565 | 0.516 |
| | Tier 2 | 0.751 | 0.742 | 0.850 | 0.847 | 0.523 | 0.490 |
| | Tier 3 | 0.731 | 0.714 | 0.876 | 0.875 | 0.539 | 0.537 |
| **Multi-class** | Tier 1 | 0.688 | 0.686 | 0.712 | 0.708 | 0.365 | 0.346 |
| | Tier 2 | 0.551 | 0.508 | 0.526 | 0.533 | 0.316 | 0.259 |
| | Tier 3 | 0.565 | 0.528 | 0.628 | 0.628 | 0.312 | 0.302 |

Table 2: Classification Performance (Binary and Multi-class) across different tiers of prompts


a photo of a person sitting
a person seated on a chair
a person seated with hips and knees bent at right angles

(a) Grad-CAM for three phrasings of the “sitting” concept, showing saliency on the chair and hip–knee region.


a photo of a person standing
a person standing still and upright, feet planted
a person standing with legs straight and torso vertical

(b) Grad-CAM for three phrasings of the “standing” concept, showing saliency on the legs and torso.

Figure 3: Grad-CAM visualizations for different phrasings of the concepts “sitting” and “standing”. Increased specificity in phrasing leads to more focused saliency in relevant body regions.

when such constraints are used, and decreased reliance on background texture or incidental objects. The benefit is class dependent. Categories that are well captured by a nominal phrase, such as sitting in uncluttered scenes, receive limited additional gain. Categories that are visually adjacent in a still image, such as standing versus walking or running, benefit from geometric prompts because these encode spatial structure that separates the classes without introducing scene bias.

These observations support a simple policy for low-resource use. Prefer label-style prompts as the default in closed-set classification with pre-trained encoders. Introduce compact geometric descriptors selectively for pairs that remain ambiguous, and verify with attribution that attention shifts from background to pose-relevant regions. Reserve action verbs for cases where the class definition truly requires dynamic semantics, since such wording is not consistently grounded in single images.

### 5.2 Comparative Model Behaviour and Calibration

Across encoders, closed-set zero-shot performance tracks the pre-training objective. OpenCLIP and MetaCLIP optimise a soft-max contrastive loss with a learned temperature, which induces competition among text candidates and yields larger similarity margins in classification. SigLIP optimises independent sigmoid scores for pairs, which favours retrieval but produces flatter score distributions in a closed set. The flatter distributions manifest as smaller top-1 minus top-2 margins and greater sensitivity to prompt phrasing, particularly for visually adjacent classes.

Calibration follows the same pattern. After unit normalisation of embeddings, a single temperature applied to cosine similarities brings CLIP-family confidences into closer agreement with accuracy. The same treatment is less effective for SigLIP model because the training objective does not enforce cross-class competition, and confidence therefore reflects pairwise affinity rather than calibrated class probability. Reliability curves and expected calibration error consequently favour OpenCLIP and MetaCLIP 2 under a shared temperature, whereas SigLIP remains comparatively miscalibrated or require tier-specific scaling.

Baselines clarify the role of alignment and structure. DINOv3 and a standard ViT combined with frozen sentence embeddings underperform and calibrate poorly because the image and text spaces are learned independently rather than jointly. YOLOv11-Pose with a simple geometric decision rule is competitive when keypoints are detected with confidence, which indicates that explicit pose structure can substitute for language supervision when the detector is reliable. Taken together, these observations suggest that, in data-scarce image-based recognition, cross-modal alignment with a contrastive objective provides stronger closed-set behaviour, while geometric priors provide a complementary path when alignment is weak or text supervision is constrained.

### 5.3 Language-Free and Pose-Based Baselines under Data Scarcity

Vision-only encoders such as DINOv3 and a standard ViT provide a language-free reference that isolates the value of cross-modal alignment. When image embeddings are compared to sentence embeddings from an unrelated text model, the spaces are not jointly learned. As a result the cosine geometry reflects two independent objectives rather than class evidence. This mismatch explains the weaker separability and the poor calibration that appear even when preprocessing is held constant. The baselines are therefore informative as a lower bound. They confirm that generic visual features carry some signal for posture, yet they also show that alignment with text during pre-training is the primary driver of robust zero-shot classification.

A pose-based baseline introduces a different kind of supervision that is structural rather than linguistic. YOLOv11-Pose produces 2D keypoints, and a deterministic rule maps joint configuration to the three classes. When detections are confident, the rules are competitive because they test explicit geometric relations that are stable in a single frame. However, performance depends on detection coverage. Occlusion, truncation, unusual viewpoints, and multiple persons reduce keypoint quality and lead to abstentions or incorrect geometry, which directly lowers accuracy. Reporting coverage alongside accuracy is therefore necessary. On the covered subset the baseline demonstrates that posture can be resolved without any text, while the uncovered subset clarifies where structural priors fail.

These baselines contribute two practical insights for low-resource use. First, if language supervision is restricted because of privacy or deployment constraints, a pose pipeline can recover a substantial fraction of performance provided that person detection is dependable. Second, if a language-free heuristic is required for simplicity, cosine scoring between DINOv3 or ViT features and frozen sentence embeddings should be treated as a diagnostic tool rather than as a calibrated classifier. In combination with the multimodal results, the baselines indicate that cross-modal alignment should be the default, and that explicit pose structure is a useful fallback when alignment is unavailable or when prompts cannot be used.

### 5.4 Attention Maps and Error Patterns in Still-Image HAR

Attribution on the sitting examples shows a consistent shift as prompt specificity increases. The label-style prompt yields broad responses that cover the person and nearby objects. Adding an action cue narrows the response toward the pelvis and the supporting surface. Geometric phrasing concentrates the map on hips, knees, and the contact region with the chair. This progression indicates that geometric wording encourages the model to prefer pose evidence over contextual cues.

For standing, the label-style prompt again produces diffuse maps with noticeable activation on salient background regions as seen in Figure 3b. The action cue that mentions stillness reduces spread and increases activation around the legs and feet. The geometric formulation further localises energy along the vertical axis of the body, especially the shins and torso. When predictions are incorrect for standing, the maps typically remain broad and include background structure, which suggests insufficient reliance on limb configuration in those cases.

Two practical uses follow. First, attribution can serve as a prompt diagnostic: adopt geometric phrasing when maps remain diffuse under a label-style prompt, and retain the minimal prompt when maps are already concentrated on limbs and joints. Second, report simple map statistics alongside accuracy, such as the proportion of normalised heat inside a person region and the entropy of the map. Higher in-person proportion and lower entropy correlate with the tighter, pose-focused responses observed for the geometric prompts in both sitting and standing.

### 5.5 Practical Implications, Robustness, and Limitations

**Practical implications** In data-scarce settings, a label-style prompt for each class with unit-normalised embeddings and a single temperature applied to cosine scores is a strong baseline. When confusions remain for visually adjacent categories, replace the label with a compact geometric description for those specific classes. Monitor decision confidence with the top-one minus top-two similarity margin. Introduce an abstention rule based on a margin threshold for low-confidence cases. If language supervision is not available, a pose pipeline that uses YOLOv11-Pose with a deterministic geometric rule provides an alternative, provided that keypoint detection is reliable.

**Robustness considerations** Performance depends on image framing and resolution. Crops that remove feet or hips reduce margins for posture classes, therefore detection and resizing should preserve the lower body. Resolution influences prompts that encode limb configuration. The native 224 input supports fair comparison, while higher resolution can improve separation when resources allow. Paraphrases within a tier can shift scores, so a small prompt ensemble per class stabilises predictions with limited overhead. Calibration differs across encoders. Fit a single temperature once per model and keep it fixed across tiers to preserve comparability. For the pose baseline, report coverage since occlusion, truncation, and small subjects reduce the fraction of usable detections.

**Limitations** Utilizing a small COCO-derived subset (285 images) focusing only on three single-person posture classes (sitting, standing, walking/running) from still images, inherently limits its external validity and applicability to real-world Human Activity Recognition (HAR) where temporal cues are essential. Additionally, the discovery of "prompt overfitting" is based solely on handcrafted English prompts across a limited set of encoders (OpenCLIP, MetaCLIP 2, SigLip) and verified using only Grad-CAM for qualitative analysis. This means the conclusions apply most directly to static, image-based posture classification under severe data scarcity.

## 6 Conclusion

This study examined how natural language prompt specificity functions as a form of supervision at inference time for zero shot posture recognition under extreme data scarcity. By holding all visual processing constant and varying only the wording of class prompts across three tiers, we identified a consistent and counterintuitive pattern in modern VLMs: stronger encoders such as MetaCLIP 2 and OpenCLIP perform best with minimal label-style prompts, while additional descriptive detail reduces accuracy, a phenomenon we term *prompt overfitting*. Conversely, lower-performing models benefit from compact geometric descriptions, particularly for visually adjacent postures. Attribution analyses further showed that prompt wording shifts model attention toward or away from pose relevant regions, clarifying why specificity can help or hinder depending on encoder strength.

Together with comparisons to vision only and pose based baselines, these findings provide practical guidelines for deploying VLMs in low-resource, image-based Human Activity Recognition (HAR) settings. Label-style prompts serve as a strong default, while geometric descriptions can assist when distinctions are subtle and model capacity is limited. More broadly, the results demonstrate that prompt semantics act as implicit priors that can either reinforce or misalign the visual grounding of multimodal encoders.

Future work should extend this analysis to larger and more diverse datasets, multi-person and occlusion-heavy scenes, multilingual and LLM-generated prompts, additional VLM backbones, and more advanced attribution techniques. Such work is necessary to fully understand and generalise the behavioural patterns observed in this study.

## References

Mahmoud Ali, Di Yang, and Franccois Br’emond. 2024. Are visual-language models effective in action recognition? a comparative study. *ArXiv*, abs/2410.17149.

Massimo Bosetti, Shibingfeng Zhang, Bendetta Liberatori, Giacomo Zara, Elisa Ricci, and Paolo Rota.

2024. Text-enhanced zero-shot action recognition: A training-free approach. *ArXiv*, abs/2408.16412.

Lincan Cai, Jingxuan Kang, Shuang Li, Wenxuan Ma, Binhui Xie, Zhida Qin, and Jian Liang. 2025. From local details to global context: Advancing vision-language models with attention-based selection. *ArXiv*, abs/2505.13233.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. 2025. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2501.01234*.

Pablo Meseguer Esbri, Rocío del Amor, and Valery Naranjo. 2024. Mi-visionshot: Few-shot adaptation of vision-language models for slide-level classification of histopathological images. *ArXiv*, abs/2410.15881.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. 2025a. pfedmma: Personalized federated fine-tuning with multi-modal adapter for vision-language models. *ArXiv*, abs/2507.05394.

Sajjad Ghiasvand, Haniyeh Ehsani Oskouie, Mahnoosh Alizadeh, and Ramtin Pedarsani. 2025b. Few-shot adversarial low-rank fine-tuning of vision-language models. *ArXiv*, abs/2505.15130.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, Deva Ramanan, Piotr Dollár, and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D. Bagdanov. 2024. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *European Conference on Computer Vision*.

Sarah Pratt, Rosanne Liu, and Ali Farhadi. 2022. What does a platypus look like? generating customized prompts for zero-shot image classification. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Hao Shao, Shengju Qian, and Yu Liu. 2020. Temporal interlacing network. *ArXiv*, abs/2001.06499.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. 2025. Dinov3.

Sreekanth. 2024. Sreekanth3096/vit-coco-image-classification · hugging face. *Sreekanth3096/vit-coco-image-classification · Hugging Face*.

Illia Volkov, Nikita Kisel, Klára Janousková, and Jirí Matas. 2025. Image recognition with vision and language embeddings of vlms.

Qian-Wei Wang, Yuqiu Xie, Letian Zhang, Zimo Liu, and Shu-Tao Xia. 2025. Pre-trained vision-language models assisted noisy partial label learning. *ArXiv*, abs/2506.03229.

Ming-Kuan Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, Xiaoshuai Sun, and Rongrong Ji. 2022. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *AAAI Conference on Artificial Intelligence*.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and X. Wang. 2022. Groupvit: Semantic segmentation emerges from text supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123.

Zhenlin Xu, Yi Zhu, Tiffany Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. 2023. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1827–1836.

Rowan Zellers and Yejin Choi. 2017. Zero-shot activity recognition with verb attribute induction. *ArXiv*, abs/1707.09468.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:5625–5644.

Yabin Zhang, Wen-Qing Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. 2024. Dual memory networks: A versatile adaptation approach for vision-language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28718–28728.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krahenbuhl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. *ArXiv*, abs/2201.02605.

# BengaliFig: A Low-Resource Challenge for Figurative and Culturally Grounded Reasoning in Bengali

**Abdullah Al Sefat**
Independent Researcher
abd.al.sefat@gmail.com

## Abstract

Large language models excel on broad multilingual benchmarks but remains to be evaluated extensively in figurative and culturally grounded reasoning, especially in low-resource context. We present **BengaliFig**, a compact yet richly annotated challenge set that targets this gap in Bengali, a widely spoken low-resourced language. The dataset contains 435 unique riddles drawn from Bengali oral and literary traditions. Each item is annotated along five orthogonal dimensions capturing reasoning type, trap type, cultural depth, answer category, and difficulty, and is automatically converted to multiple–choice format through a constraint–aware, AI–assisted pipeline. We evaluate eight frontier LLMs from major providers under zero–shot and few–shot chain–of–thought prompting revealing consistent weaknesses in metaphorical and culturally specific reasoning. BengaliFig thus contributes both a diagnostic probe for evaluating LLM robustness in low-resource cultural contexts and a step toward inclusive and heritage-aware NLP evaluation. Data and evaluation code is available at `https://github.com/chaoSefat/Bengali-Fig`

## 1 Introduction

Over the years we have seen several largescale Question-Answer(QA) datasets such as SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017) and Natural Questions (Rajpurkar et al., 2016). Datasets such as DROP (Dua et al., 2019), ARC (Clark et al., 2018) and MMLU (Hendrycks et al., 2021) push models towards deeper knowledge and structured reasoning skills rather than simple literal QA. Large Language Models (LLMs) have achieved impressive results on such large-scale benchmarks. However, figurative, metaphorical and culturally grounded reasoning are blindspots of these large scale datasets. While some work has been done in metaphor detection (Leong et al., 2020), (Maudslay et al., 2020), (Lu and Wang, 2017), (Wang et al., 2019),the focus is on high-resourced languages such as Chinese and English, leaving many widely spoken but under-resourced languages unexplored.

Small, focused probe tasks have proven useful for diagnosing specific reasoning capabilities (e.g., the Winograd Schema Challenge (Levesque et al., 2012), HANS (McCoy et al., 2019), StressTest (Naik et al., 2018)). Such resources demonstrate that fewer deliberately curated examples can reveal failure modes that large corpora and benchmarks do not reveal. This is particularly important for low-resource languages, where cultural and oral traditions encode figurative reasoning that is rarely captured by existing datasets.

Riddles are an oral and literary form rich in metaphor, misdirection, and local knowledge form a natural diagnostic arena but are absent from current evaluation suites. Bengali figurative riddles often encode perceptual and symbolic cues, referencing color, form, sound, and motion and thus offering a textual lens into reasoning that naturally spans multiple modalities. Bengali is the $7^{th}$ most spoken language in the world[1], yet no evaluation specifically probes figurative or culturally grounded reasoning in Bengali.

To address this gap we present **BengaliFig**, a challenge set crafted to stress-test figurative reasoning and cultural grounding in Bengali. Our contributions are threefold:

1. **Challenge set creation:** We curate and release a corpus of 435 unique Bengali riddles, each manually deduplicated, normalized and structured as Multiple Choice Question (MCQ) format.

2. **Multi-axis Annotation:** We annotate our curated QA dataset over five orthogonal di-

---

[1] `https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/`

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 58–77
December 23, 2025 ©2025 Association for Computational Linguistics

mensions capturing cognitive and cultural attributes.

3. **Comprehensive evaluation:** We probe eight frontier LLMs under zero-shot and few-shot chain-of-thought prompting. We then analyze their performance breakdown over the annotated dimensions and prompting techniques.

Our results demonstrate that majority of the frontier LLMs struggle significantly with Bengali riddles. BengaliFig thus fills a critical gap by providing a culturally grounded, low-resource testbed for probing LLM robustness and for guiding more inclusive NLP research.

## 2 Related Works

Multilingual benchmarks such as FLORES-200 (Team et al., 2022), XTREME (Hu et al., 2020), and IndicGLUE (Kakwani et al., 2020) include Bengali but focus primarily on translation, classification, or factual QA. Dedicated Bengali resources include BanglaNLG for natural language generation (Bhattacharjee et al., 2023), BanglaRQA for reading comprehension (Ekram et al., 2022), Vashantor for dialect translation (Faria et al., 2023), BenNumEval for numerical reasoning (Ahmed et al., 2025), and BEnQA for middle- and high-school QA (Shafayat et al., 2024). These tasks remain largely literal and do not assess figurative, metaphorical, or culturally embedded reasoning.

Research on figurative language has focused primarily on high-resource languages such as English and Chinese. Prior work includes metaphor detection (Leong et al., 2020; Maudslay et al., 2020; Lu and Wang, 2017; Wang et al., 2019) and broader figurative understanding (Jang et al., 2023; Lai and Nissim, 2024). Riddle-focused resources such as BiRdQA (Zhang and Wan, 2021), CC-Riddles (Xu et al., 2023), and Visual Riddles (Bitton-Guetta et al., 2024) probe models' ability to integrate metaphor, ambiguity, and cultural knowledge. However, these datasets remain concentrated in high-resource languages and do not extend to Bengali.

Despite Bengali being one of the world's most widely spoken languages, no benchmark targets metaphorical, figurative, or culturally grounded reasoning. Such abilities are deeply rooted in cultural context, making Bengali riddles a natural stress-test for LLMs. Carefully constructed, high-signal examples can reveal failure modes invisible to large benchmarks (Levesque et al., 2012; McCoy et al., 2019; Naik et al., 2018), motivating our probe-set design.

## 3 BengaliFig

We describe our methodology for BengaliFig dataset construction in this sections. Our steps are described below:

### 3.1 Data Collection and Preprocessing

We built the BengaliFig corpus by scraping riddles from blogs, forums, and digital archives, then filtering and cleaning them through a compact three–stage pipeline: *deduplication*, *normalization*, and a final *manual audit*.

### 3.1.1 Deduplication

To ensure that every item is unique yet representative, we combined automatic retrieval with human checks. Each riddle question is a Unicode string $q_i$. For every pair $(q_i, q_j)$ we compute the normalized Levenshtein distance

$$d(q_i, q_j) = \frac{\mathrm{lev}(q_i, q_j)}{\max(|q_i|, |q_j|)} \in [0, 1], \qquad (1)$$

where $\mathrm{lev}(\cdot, \cdot)$ is the minimal edit count. Pairs with $d(q_i, q_j) \le \tau$ were flagged as candidates, starting with $\tau = 0.10$ for high precision and gradually relaxed to $0.30$ for recall. Flagged pair was automatically deduplicated only if the answers were perfect overlaps. Within each candidate cluster, we kept the element with smallest identifier as the canonical form. Native speakers then reviewed all the remaining candidates to discard duplicates. This hybrid design delivered near–perfect precision while capturing subtle paraphrases. Our initial Collection consisted of 770 entries. 238 were removed after the automatic and manual deduplication.

### 3.1.2 Answer Normalization

We standardized answer text by removing extraneous punctuation and isolating the core answer when sources contained extra explanation. For a raw answer $\alpha$ and delimiter set $\mathcal{S} = \{$“:”, “–”, “|”, “—”$\}$ we define, $\hat{\alpha} = \mathrm{first\_split}(\alpha, \mathcal{S})$, logging them $\mathcal{E} = \{(\alpha, \hat{\alpha})\}$ for manual audits.

### 3.1.3 Manual Audit

Finally, two native speaker auditors performed a full pass to catch residual issues with answer normalization edits, mistranslations, malformed riddles, hidden duplicates, or question–answer mismatches. Unsalvageable entries were removed; ambiguous

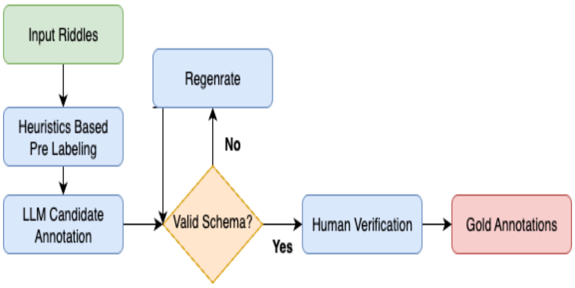


Figure 1: LLM-assisted annotation pipeline. Heuristic priors $p_0$ seed LLM predictions $\hat{y}$, validated against label set $\mathcal{Y}$ and finalized as $y^*$ by human annotators.

but valuable riddles were lightly edited to maintain fluency and logical consistency. Total of 97 entries were removed after manual audit.

### 3.2 LLM-Assisted Human Annotation

Annotating riddles is challenging because solutions hinge on culture, figurative language, and multi-step reasoning. Each riddle is labeled along five orthogonal dimensions, making purely manual work costly. We therefore adopt an **LLM-assisted framework** in which a large language model proposes candidate labels that are then verified and, if necessary, corrected by human annotators. We describe our annotation schema below where we just list down the five orthographic dimensions and their set of possible values. In Appendix A.7 we provide the annotation schema with detailed explanation of the labels alongside the LLM prompt.

**Annotation Schema.** The five dimensions capture complementary cognitive and cultural properties:

**Reasoning Type**

($r \in \mathcal{R}$): { *metaphorical*, *commonsense*, *descriptive*, *wordplay*, *logical_deduction*, *compound* }

**Trap Type**

($t \in \mathcal{T}$): *surface_literal*, *multiple_valid*, *culturally_specific*, *linguistic_trick*, *misdirection*, *archaic_reference*, *none*.

**Cultural Depth**

$(c \in \mathcal{C}) : \{\text{universal}, \text{cultural\_specific}\}$.

**Answer Type**

($a \in \mathcal{A}$): *place*, *person*, *animal*, *plant*, *object*, *natural_phenomenon*, *body_part*, *food_drink*, *concept*, *quantity*, *text_symbol*.

**Difficulty**

$(d \in \mathcal{D}) : \{\text{easy}, \text{medium}, \text{hard}\}$.

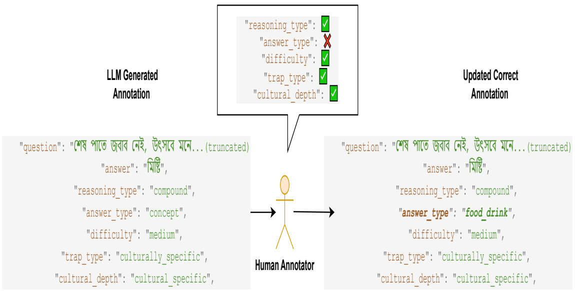
All sets $\mathcal{R}, \mathcal{C}, \mathcal{T}, \mathcal{A}, \mathcal{D}$ are mutually exclusive and exhaustive.

**Framework:** The annotation pipeline, illustrated in Figure 1, proceeds in three stages described below.

**Stage 1: Heuristic Based Pre-Labeling.** Given a riddle–answer pair $(q, \alpha)$, we first compute a vector of heuristic priors $\mathbf{p}_0 \in [0, 1]^{|\mathcal{A}|}$ for the *answer_type* label using regex patterns and gazetteer look-ups derived from Bengali morphology. For example, if $\alpha$ contains suffixes like “পুর/নগর” (`pur/nagar`) or matches any token in the lexicon set form places, $\mathcal{L}_{\text{place}}$, we set $\mathbf{p}_0[\text{place}] = 1$. Similar detectors exist for animals, plants, body parts, natural phenomena, etc. These lightweight priors injected into the prompt to stabilize the LLM generation.

**Stage 2: LLM Candidate Annotation.** Let the complete label space be $\mathcal{Y} = \mathcal{R} \times \mathcal{A} \times \mathcal{D} \times \mathcal{T} \times \mathcal{C}$, covering reasoning type, answer type, difficulty, trap type, and cultural depth. The LLM (DeepSeek V3) receives $(q, \alpha, \mathbf{p}_0)$ and a compact schema prompt, and must output a candidate annotation for each $(q, \alpha)$ pair as a single valid tuple in strict JSON format: $\hat{y} = (r, a, d, t, c) \in \mathcal{Y}$. Temperature is fixed at $\tau = 0.1$ to minimize randomness. A validator enforces type constraints; any invalid $\hat{y}$ triggers re-prompting with the same $\mathbf{p}_0$. Cost effective inference API was the deciding factor in choosing the DeepSeek for suggesting annotation. LLM's task is not to provide final annotation but suggestions in structured JSON schema which is is easy to for annotators to edit and provide the final annotation, saving time.

**Stage 3: Human Verification** Two native-speaker annotators receive a set of entries to annotate. We first test inter annotator agreement on a 5% (22 out of 435 items) stratified set, and obtained Krippendorff's alpha = 0.9034. (Comprehensive calculation in Appendix A.1). The obtained score is well above the acceptable threshold (0.85) to continue. The remaining 413 riddles were single-annotated after establishing sufficient agreement. The annotators inspect each candidate annotation $\hat{y}$ and either accept it or supply a corrected gold label $y^*$, producing the final gold-standard labels as illustrated in Figure 2. This hybrid design substantially reduces annotation effort while retaining reliability. We also observed that annotation time reduced from $\approx$ 7.3 minutes (manual) to 2.4 minutes per riddle. This human-in-the-loop design preserves cultural fidelity while reducing average annotation time.



```
"reasoning_type": ✓
"answer_type": ✗
"difficulty": ✓
"trap_type": ✓
"cultural_depth": ✓
```

```
"question": "শেষ পাতে জবাব নেই, উৎসবে মনে...(truncated)
"answer": "মিষ্টি",
"reasoning_type": "compound",
"answer_type": "concept",
"difficulty": "medium",
"trap_type": "culturally_specific",
"cultural_depth": "cultural_specific",
```

```
"question": "শেষ পাতে জবাব নেই, উৎসবে মনে...(truncated)
"answer": "মিষ্টি",
"reasoning_type": "compound",
"answer_type": "food_drink",
"difficulty": "medium",
"trap_type": "culturally_specific",
"cultural_depth": "cultural_specific",
```

Figure 2: Example of Human annotator providing final validation to LLM generated candidate annotations

### 3.3 Exploratory Data Analysis

Figure 4 summarizes the distribution of the five annotation dimensions across all 435 riddles. **Reasoning type** is dominated by *metaphorical* riddles (224, 52%), which further underscores the contribution of our dataset providing a test bench for Bengali figurative and metaphorical reasoning. **Answer types** are diverse: tangible *objects* leading (127) while culturally salient categories are also well represented. Difficulty skews toward the middle, with *medium* items forming the majority (235), only 8 questions rated *hard*, and the rest *easy*. The riddle style is largely *surface–literal* in its **trap type** (358), with smaller pockets of *linguistic_trick* (57). Majority riddles require **cultural knowledge**, with *cultural_specific* depth accounting for 285 instances.

In Figure 3 *surface–literal* traps occur across both cultural depths but are strongly concentrated in *cultural_specific* items (213 vs. 145), whereas *linguistic_trick* riddles are almost exclusively cultural (56 of 57), highlighting that deceptive wordplay is closely tied to Bengali linguistic nuance. We also observe that riddles requiring wordplay (75 of 83) and compound (63 of 72 ) reasoning tend to be culturally specific. While commonsense reasoning (31 of 42) is more universal.

Cross–label analyses (Figure 5) reveals that reasoning complexity correlates with difficulty: over half of *compound* riddles are *medium* and a notable 7 are the only cluster of *hard* questions. Metaphorical riddles tend to be more inclined towards medium difficulty, whereas *commonsense* riddles remain predominantly *easy*. The figure on the right illustrates that culturally specific riddle tends to be more difficult than the ones that can be solved with universal basic knowledge.

### 3.4 MCQ Format Creation

Drawing inspiration from (Zellers et al., 2019) we leverage LLMs to generate distractors and to create multiple-choice questions (MCQs). Using a fully automated, two-stage AI pipeline designed to balance *diversity* in candidate distractors with *precision* in final selection. The pipeline consists of 3 steps:

**Step 1: Constraint extraction and prompt conditioning.** Many Bengali riddles state explicit surface clues such as required grapheme count (অক্ষর *akṣar* 'grapheme') or properties like size, color, number, shape, or time. We apply rule-based detectors that (a) identify Bengali numeral words and Bengali digits (১, ২, ৩) when answer's required grapheme count is mentioned, and (b) flag other additional cues such as size, color, shape, count, time etc. These constraints are packed into a structured prompt and attached to each riddle together with its question, answer, reasoning type ($r \in \mathcal{R}$), and answer type ($a \in \mathcal{A}$).

**Step 2: Constraint and misdirection-aware generation.** We give the prompt with extracted constraints to a generator LLM. The generator produces $n = 6$ distractors that exploit the riddle's surface misdirection rather than copying the correct answer. Candidates must: (i) seem plausible under the surface meaning, (ii) sound natural to Bengali speakers, and (iii) follow all constraints like grapheme length. We use higher sampling temperature for diverse outputs.

**Step 3: Automated selection under explicit criteria.** A separate selector LLM ranks candidates using five criteria: misdirection power, first-instinct appeal, surface-logic coherence, constraint compliance, and diversity of traps. The selector uses lower temperature for stable results. We apply basic checks, shuffle options, and record the correct answer's position.

**Model heterogeneity:** We use two different models in our pipeline for practical and methodological reasons. DeepSeek-V3 serves as the generator, while GPT-4 handles the selection task. This division separates the generation and evaluation processes to reduce self-endorsement bias. The choice of DeepSeek for generation was driven by cost considerations and its accessible API. For the selection stage, we chose GPT-4 due to its established reliability in evaluation tasks and consistent

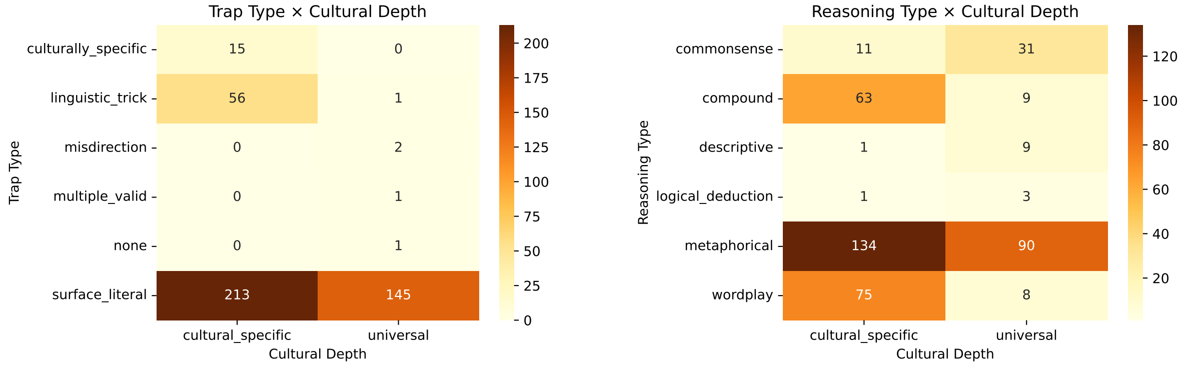


Figure 3: Cross–label relationships: Trap type vs. cultural depth(left) and Reasoning vs. cultural depth(right)

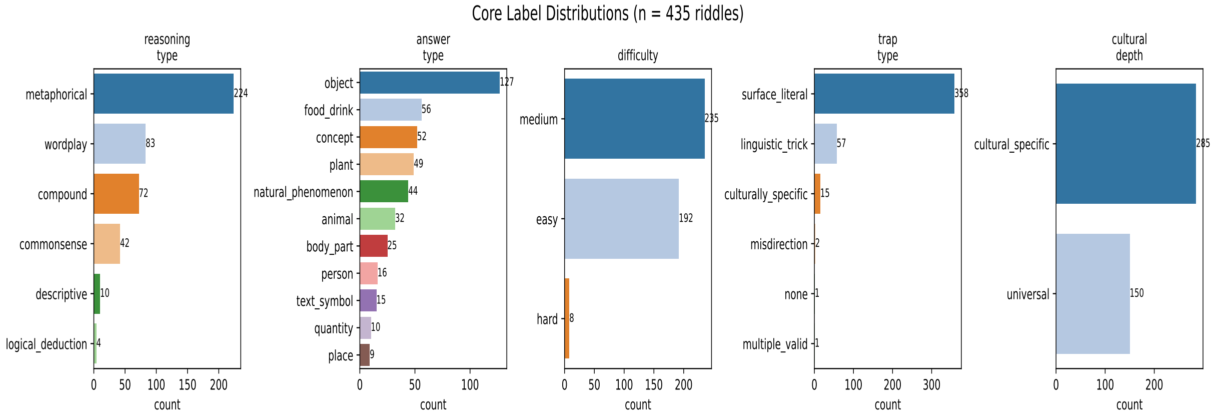


Figure 4: Core label distributions across the five annotation dimensions.

performance across different tasks. (Full prompts in Appendix A.2)

**Observations** In Bengali graphemes represent syllables. Even with constraint blocks, generated distractors often ignored grapheme count requirement. Not a single riddle with a grapheme constraint received a complete set of options meeting that limit. The failure of LLMs to generate distractor options conforming to the grapheme count reveals their graphemic and phonological weakness in non Latin scripts. In contrast, answer-type constraints (e.g., country, fruit, language) were largely respected.

## 4 Experiments and Results

To rigorously assess the figurative reasoning capabilities of large language models (LLMs) on BengaliFig, we developed a comprehensive and robust evaluation framework. This framework is designed to handle multiple model providers, support diverse evaluation modes (zero-shot, and few shot chain-of-thought prompting (CoT)), and guarantee reproducibility through systematic result logging and metadata tracking.

We evaluate a diverse set of LLMs across major providers: (i) **OpenAI**: GPT-4.1 and GPT-5, (ii)**Anthropic**: Claude Sonnet 4.0 and Claude Opus 4.1, (iii)**DeepSeek**: DeepSeek-V3.1[2], (iv)**Meta**: LLaMA-4 Maverick, LLaMA-4 Scout, (v)**Qwen**: Qwen3-235B

### 4.1 Zero-Shot Evaluation

We first assess all models in a strict zero-shot setting, where each riddle is presented with four multiple-choice options and models must return only the single correct letter (A-D). Accuracy is reported over the entire 435-item test set and key annotation dimensions. See Appendix A.3 for prompt and result reproducibility.

**Overall Performance Rankings.** Table 1 presents the comprehensive performance hierarchy. **GPT-5 achieves the highest accuracy at 82.3%**, followed closely by **Claude-Opus-4.1 at 79.8%**, establishing a clear top tier. Performance then

---

[2] DeepSeek-V3.1 was used for final evaluation, upgrading from V3 used in earlier steps.

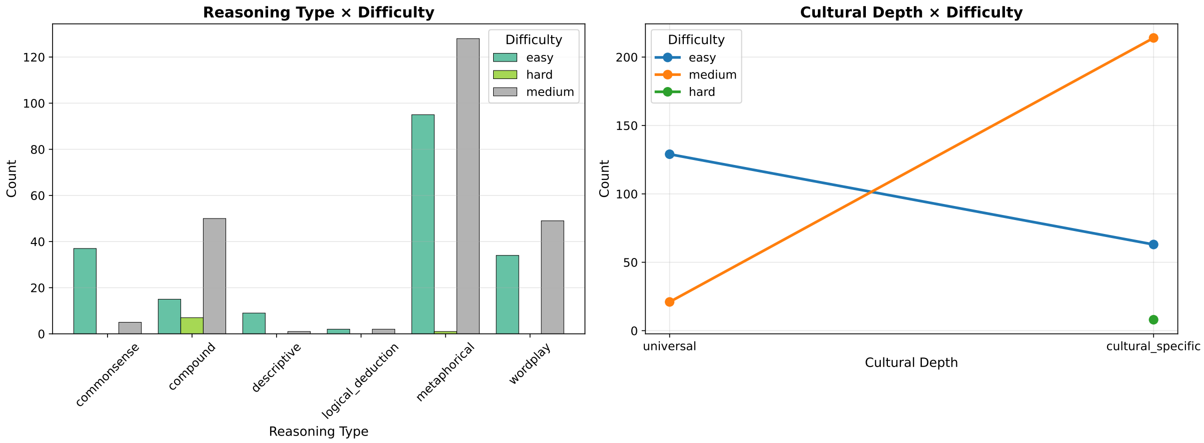


Figure 5: Cross–label relationships: reasoning type vs. difficulty (left); and cultural depth vs. difficulty (right)

| Model | Overall Acc. |
|---|---|
| GPT-5 | **82.3** |
| Claude-Opus-4.1 | 79.8 |
| GPT-4.1 | 69.0 |
| LLaMA-4 Maverick | 63.2 |
| DeepSeek-V3.1 | 59.8 |
| Qwen3-235B | 58.6 |
| LLaMA-4 Scout | 55.2 |
| Claude-Sonnet-4.0 | 50.8 |

Table 1: Zero-shot overall accuracy (%)

drops substantially to GPT-4.1 (69.0%), creating a notable 10.8-point gap that suggests qualitative differences in reasoning capabilities. The remaining models cluster in the 55-63% range, with Claude-Sonnet-4.0 performing weakest at 50.8% and barely exceeds random chance in our 4-option multiple-choice format.

**Reasoning Type Breakdown.** Figure 6 reveals pronounced variation in accuracy across reasoning categories. All models perform well on *descriptive* and *logical deduction* tasks, with top performers achieving perfect accuracy (100%) while, *metaphorical reasoning* poses was more challenging. Even for leading models such as GPT-5 and Claude-Opus plateau around 80-81%, suggesting inherent difficulty in abstract conceptual mapping within the Bengali cultural context. *Wordplay* emerges as the most discriminative category, where performance gaps exceed 40 percentage points. GPT-5 leads at 84.3%, while Claude-Sonnet-4.0 achieves only 39.8%. This significant gap underscores the linguistic sophistication required for Bengali phonetic and orthographic manipulation, where models must simultaneously process sound patterns, semantic ambiguity, and cultural references. *Commonsense* and *compound* reasoning showed intermediate difficulty levels, with top models reaching 81-83% accuracy.

**Trap Type.** Surface-literal traps dominate the dataset (358 of 435), so any apparent correlation between trap susceptibility and overall accuracy may be confounded by the class imbalance; detailed analysis is provided in Appendix A.6. Our analysis of trap-type correlations is only exploratory.

**Difficulty and Cultural Depth Analysis.** Tables 2 and 3 reveal systematic performance patterns across BengaliFig's annotation dimensions. Difficulty levels show clear stratification: accuracy decreases monotonically from Easy (Mean: 70.1%, Range: 47.4–85.9%) to Medium (61.7%, 53.2–80.0%) to Hard (29.7%, 0.0–62.5%). The substantial 40.4-point mean gap between Easy and Hard categories validates our annotation scheme while demonstrating genuine cognitive challenges. Even GPT-5 achieves only 62.5% on Hard riddles. Cultural depth analysis reveals consistent but more subtle effects: universal riddles outperform cultural-specific counterparts across all models, with a mean advantage of 10.0 percentage points. This systematic disparity (ranging from +5.6 for GPT-5 to +21.7 for GPT-4.1) indicates that cultural knowledge requirements impose additional cognitive load beyond linguistic competence alone. Notably, the cultural gap persists even for extensively multilingual models, suggesting deeper pragmatic understanding challenges rather than surface-level cultural fact retrieval limitations.

**Grapheme–Constraint Evaluation.** Some riddles explicitly specify that the correct answer must contain a fixed number of Bengali graphemes, a cue that humans can easily exploit to eliminate

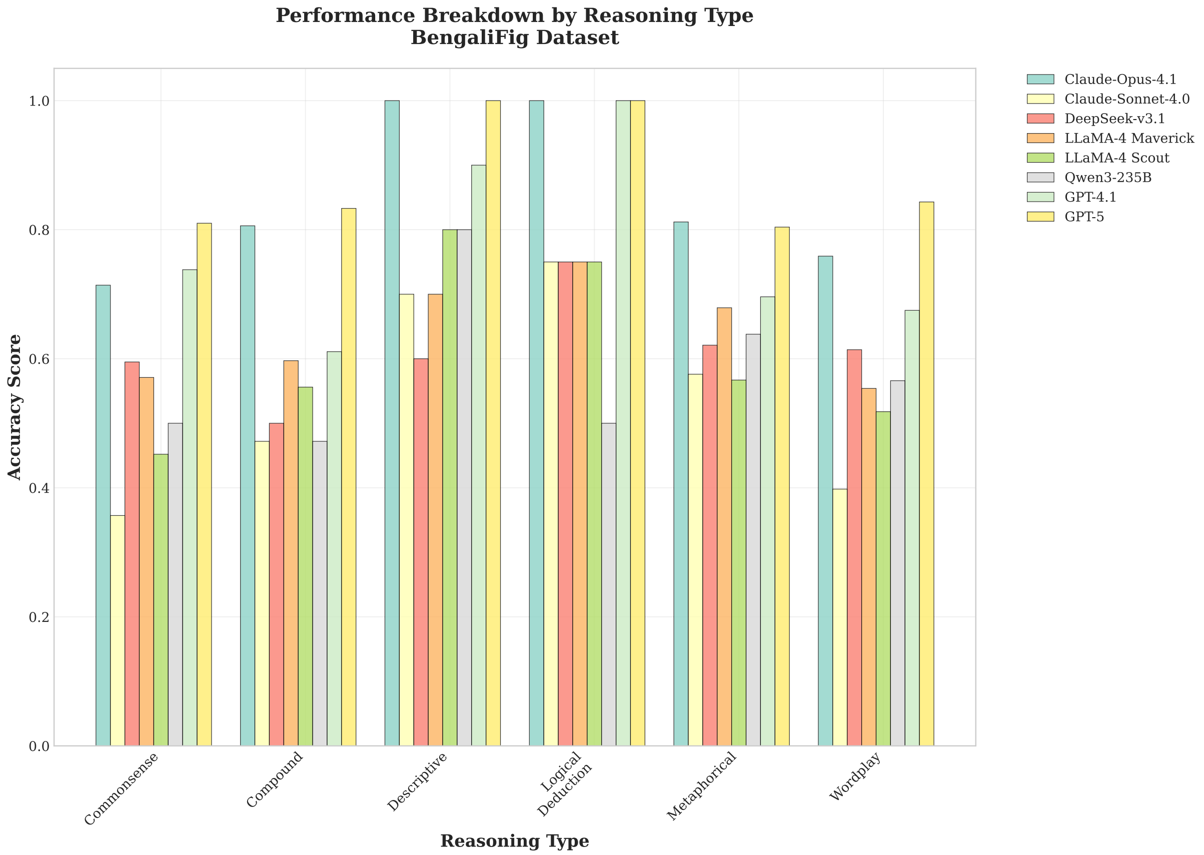


Figure 6: Performance breakdown by reasoning type, revealing significant variation across cognitive categories.

implausible options. Despite explicitly prompting this constraint, none of the LLMs produced a full set of distractors adhering to the grapheme counts, effectively turning these riddles into elimination tasks for humans. Across this 28–item subset, GPT-5 achieved the highest accuracy (85.7%), followed by GPT-4.1 (60.7%) and Claude-Opus-4.1 (42.9%). All other models, including Claude-Sonnet-4.0 (28.6%), Qwen3 (32.1%), DeepSeek-V3.1 (35.7%), and LLaMA-4 variants (7.1–42.9%) performed much worse. This sharp degradation suggests that current LLMs, further reinforces that even strong multilingual ones, struggle to interpret or consistently respect orthographic and phonological constraints in non-Latin scripts.

**Proprietary API Cost.** Considering API usage cost in proprietary models, despite leading in accuracy, GPT-5 and Claude-4.1-Opus incur several-fold higher API costs than GPT-4.1(Appendix A.4).

| Model | Easy | Med | Hard |
|---|---|---|---|
| GPT-5 | 85.9 | 80.0 | 62.5 |
| Claude-Opus-4.1 | 84.4 | 77.0 | 50.0 |
| GPT-4.1 | 78.6 | 62.6 | 25.0 |
| LLaMA-4 Maverick | 69.3 | 58.7 | 50.0 |
| DeepSeek-V3.1 | 69.8 | 53.2 | 12.5 |
| Qwen3-235B | 66.7 | 54.0 | 0.0 |
| LLaMA-4 Scout | 58.9 | 54.0 | 0.0 |
| Claude-Sonnet-4.0 | 47.4 | 54.0 | 37.5 |
| **Mean** | **70.1** | **61.7** | **29.7** |
| **Std Dev** | 13.7 | 10.1 | 24.1 |

Table 2: Zero-shot accuracy (%) by difficulty level on BengaliFig.

## 4.2 Few-Shot Chain-of-Thought Analysis

To investigate whether structured reasoning can improve performance on challenging riddles, we conduct few-shot Chain-of-Thought (CoT) evaluation on a strategically selected subset of BengaliFig. We identify the subset of ”hardest yet solvable” instances which are riddles where exactly one model succeeded in zero-shot evaluation while all others failed. This ensured that our analysis focuses on genuinely difficult but not impossible reasoning challenges. See Appendix A.5 for prompt and reproducibility notes.

**Experimental Design.** Our few-shot CoT prompt provides two Bengali riddle exemplars with explicit reasoning traces, followed by a structured three-

| Model | Cultural | Universal |
|---|---|---|
| GPT-5 | 80.4 | 86.0 |
| Claude-Opus-4.1 | 78.6 | 82.0 |
| GPT-4.1 | 63.5 | 79.3 |
| LLaMA-4 Maverick | 57.5 | 74.0 |
| DeepSeek-V3.1 | 54.0 | 70.7 |
| Qwen3-235B | 54.4 | 66.7 |
| LLaMA-4 Scout | 52.3 | 60.7 |
| Claude-Sonnet-4.0 | 50.2 | 52.0 |
| **Mean** | **61.4** | **71.4** |
| **Std Dev** | 11.4 | 11.1 |

Table 3: Zero-shot accuracy (%) by cultural depth on BengaliFig.

| Model | Zero-Shot | CoT | Gain(%) |
|---|---|---|---|
| GPT-4.1 | 0.0 | 30.0 | **+30.0** |
| Claude-Opus-4.1 | 20.0 | 43.3 | +23.3 |
| DeepSeek-V3.1 | 3.3 | 26.7 | +23.3 |
| LLaMA-4 Maverick | 6.7 | 26.7 | +20.0 |
| LLaMA-4 Scout | 3.3 | 20.0 | +16.7 |
| Qwen3-235B | 6.7 | 23.3 | +16.7 |
| Claude-Sonnet-4.0 | 16.7 | 26.7 | +10.0 |
| GPT-5 | 43.3 | 43.3 | 0.0 |

Table 4: Few-shot Chain-of-Thought performance (%) on hardest yet solvable BengaliFig subset (n=30).

step methodology: (1) riddle type identification and question analysis, (2) systematic option evaluation, and (3) logical conclusion formation. This framework encourages models to decompose complex reasoning while maintaining cultural and linguistic authenticity through native Bengali instruction.

**Differential CoT Efficacy.** Table 4 reveals striking heterogeneity in CoT responsiveness across model families. **GPT-4.1 demonstrates the most substantial improvement**, achieving a 30% accuracy gain (0% → 30%) with 9 successful corrections out of 30 initially failed cases. **Claude-Opus-4.1 and DeepSeek-Chat both achieve 23.3% improvement rates**, though from different baselines, Claude-Opus from a stronger initial position (20% → 43.3%) and DeepSeek from near-zero performance (3.3% → 26.7%).

Conversely, **GPT-5 shows zero improvement with CoT** with 0% despite starting from the highest baseline (43.3%). This counterintuitive finding suggests that GPT-5's zero-shot reasoning may already be near-optimal for this task difficulty level, with CoT providing redundant rather than complementary processing.

**Baseline Performance and CoT Ceiling Effects.** An inverse relationship emerges between zero-shot accuracy and CoT improvement. Models with low initial performance gain most, while stronger models show diminishing returns. This ceiling effect may indicate that CoT mainly helps bridge basic reasoning gaps rather than refine high-level reasoning. We caution that this finding is based on a limited 30-item subset. Across models, accuracy improvements remain below 30%, with no system exceeding 43.3%, aligning with the "hardest yet solvable" design and underscoring the cognitive difficulty of culturally grounded Bengali riddles.

**Final Insights.** Few-shot CoT results reveal that (1) structured reasoning aids mid-tier models but yields limited benefit for top-tier ones, and (2) cultural–linguistic reasoning challenges persist despite explicit reasoning cues. These patterns suggest that deeper cultural grounding, not additional prompting, is key to advancing performance on BengaliFig.

## 5 Conclusion

We introduced **BengaliFig**, a small but carefully constructed challenge set for probing figurative and culturally grounded reasoning in Bengali. Our 435 riddles are annotated along five orthogonal dimensions and converted to multiple–choice format through an AI–assisted pipeline. Evaluation of eight frontier LLMs shows that even state-of-the-art systems struggle, especially with metaphorical and culturally specific riddles. Few-shot chain-of-thought prompting yields only limited gains, confirming diminishing returns for explicit reasoning guidance. A focused analysis of riddles containing explicit Bengali grapheme-count clues reveals a further weakness: most models ignore simple phonological constraints in this non-Latin script, leading to sharp accuracy drops. These findings highlight persistent gaps in cross-lingual and script-aware reasoning and underscore the need for resources that emphasize depth and cultural specificity rather than scale. Although BengaliFig is a text-only resource, many riddles evoke inherently multimodal reasoning, linking linguistic metaphor with perceptual and sensory imagery. Future extensions could therefore explore how multimodal models grounded in language, vision, and sound handle such culturally embedded reasoning tasks in low-resource contexts. We release the data, prompts, and scripts to support future work on figurative and culturally informed evaluation in low-resourced languages.

## Limitations

Our current design focuses solely on textual reasoning, although many riddles implicitly reference visual, auditory, or tactile attributes that future multimodal extensions could capture. Our design as a focused challenge set introduces several constraints. First, the probe set is small (435 riddles), which limits its statistical power for fine-grained comparisons and cannot cover the full range of Bengali figurative language. Second, although each item is annotated along five dimensions with native-speaker verification, annotation was performed by only two annotators. After a small pilot to check inter-annotator agreement, the remaining data were split between them rather than double-annotated, so agreement estimates are limited and some subtle labels may reflect individual judgment. Third, our evaluation of few-shot chain-of-thought (CoT) prompting was restricted to a curated subset of the hardest but solvable riddles. This provided useful evidence that CoT helps mid-tier models but does not significantly raise overall reasoning ability, yet running few-shot CoT across the entire probe set could yield additional insights. Fourth, we did not obtain a human performance baseline. Although we planned a small user study to compare human solvers with LLMs, participation relied on voluntary sign-ups and we did not receive enough responses to draw meaningful conclusions.

## Ethics Statement

All riddles were collected from publicly available Bengali websites and digital archives. We removed entries containing personally identifiable information or offensive content and included only items suitable for open research release. Two native speakers performed the annotations after a small pilot to check inter-annotator agreement.

The dataset is released solely as an evaluation resource. Its small size makes it unsuitable for training large models, but it could still be misused to overstate cultural competence. We therefore document its scope and limitations and encourage responsible use in research on figurative reasoning and cross-lingual evaluation.

## References

Kawsar Ahmed, Md Osama, Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2025. BenNumEval: A benchmark to assess LLMs' numerical reasoning capabilities in Bengali. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17782–17799, Vienna, Austria. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 139561–139588. Curran Associates, Inc.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fatema Tuj Johora Faria, Mukaffi Bin Moin, Ahmed Al Wase, Mehidi Ahmmed, Md. Rabius Sani, and Tashreef Muhammad. 2023. Vashantor: A large-scale multilingual benchmark dataset for automated translation of bangla regional dialects to bangla language. *Preprint*, arXiv:2311.11142.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020.

Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Comput. Surv.*, 56(10).

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of mandarin chinese. *Language Resources and Evaluation*, 51:663–694.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 221–226, Online. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. 2024. BEnQA: A question answering benchmark for Bengali and English. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177, Bangkok, Thailand. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

{Ben Pin Yun} Wang, Xiaofei Lu, {Chan Chia} Hsu, {Eric Po Chung} Lin, and Haiyang Ai. 2019. *Linguistic metaphor identification in Chinese*, pages 247–265. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company, Netherlands.

Fan Xu, Yunxiang Zhang, and Xiaojun Wan. 2023. Ccriddle: A question answering dataset of chinese character riddles. *Preprint*, arXiv:2206.13778.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. 2021. Birdqa: A bilingual dataset for question answering on tricky riddles. *ArXiv*, abs/2109.11087.

# A Appendix

## Contents

A.1 Inter–Annotator Reliability Calculation . . . . . . . . . . . . . . . . 11
A.2 Prompt Templates Used for MCQ Format Creation . . . . . . . . . . 11
  A.2.1 Distractor Suggestion Prompt 11

A.2.2 Distractor Selection Prompt 12
A.3 Zero–Shot Evaluation Prompt and Reproduction Guide . . . . . . . . 13
A.3.1 Full Prompt Template . . . 13
A.3.2 Reproducibility Notes . . 13
A.4 Proprietary Model Evaluation Cost on Zero Shot . . . . . . . . . . . 13
A.5 Few–Shot Chain-of-Thought Prompt and Reproduction Guide . 14
A.5.1 Full Prompt Template . . . 14
A.5.2 Reproducibility Notes . . 15
A.6 Trap Type Analysis . . . . . . . . 15
A.7 LLM Assisted Riddle Annotation . 17
A.7.1 Annotation Guidelines . . 17
A.7.2 LLM Annotation Prompt . 19

## A.1 Inter–Annotator Reliability Calculation

To quantify inter–annotator agreement on the 5 % stratified audit set, we computed Krippendorff's $\alpha$ across all five annotation dimensions jointly.

**Data.** The audit set contained 22 riddles, each annotated along 5 independent dimensions, yielding $N = 22 \times 5 = 110$ annotation units. Each unit was labeled by two annotators.

**Krippendorff's $\alpha$.** For nominal data,

$$\alpha = 1 - \frac{D_o}{D_e},$$

where $D_o$ is the observed disagreement and $D_e$ is the expected disagreement under chance.

Because the five dimensions differ in category counts ($K_1 = 6,\ K_2 = 7,\ K_3 = 2,\ K_4 = 11,\ K_5 = 3$), the expected disagreement is the mean of the per–dimension maxima:

$$D_e = \frac{1}{5}\sum_{i=1}^{5}\left(1 - \frac{1}{K_i}\right) \approx 0.7532.$$

**Observed disagreement.** Across the $N = 110$ units the annotators disagreed on $d = 8$ units, so the observed disagreement is

$$D_o = \frac{d}{N} = \frac{8}{110} \approx 0.0727.$$

**Reliability.** Substituting into the formula,

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{0.0727}{0.7532} \approx \mathbf{0.9034}.$$

**Interpretation.** Following Krippendorff's guidelines ($\alpha \geq 0.80$ for reliable conclusions), the obtained $\alpha = 0.9034$ indicates **high agreement**. Therefore it is scientifically acceptable to proceed with the planned *non–overlapping* annotation of the remaining dataset.

**Per–dimension statistics.** Table A1 reports per–dimension disagreement counts and Krippendorff's $\alpha_i$ values on the 5% (22 items) stratified audit set. The per–dimension disagreements $(d_i)$ sum to the eight total disagreements reported in the main text. Expected disagreement $D_{e,i}$ for each dimension was computed under the maximal–disagreement assumption for nominal categories, $D_{e,i} = 1 - \frac{1}{K_i}$, and per–dimension reliabilities were obtained as $\alpha_i = 1 - D_{o,i}/D_{e,i}$. The joint reliability across all five dimensions is $\alpha \approx 0.90345$, consistent with the value reported in the main text and indicating high inter–annotator agreement.

## A.2 Prompt Templates Used for MCQ Format Creation

### A.2.1 Distractor Suggestion Prompt

You are an expert in Bengali riddles and psychological misdirection. Your task is to create {n} clever distractors that exploit the riddle's intended misdirection.

RIDDLE: {question}

CORRECT ANSWER: {answer}

STRATEGY: Bengali riddles work by misdirecting the reader toward an obvious but wrong interpretation. Your distractors should capitalize on this misdirection, NOT be similar to the correct answer. Focus on the main question the riddle seems to be asking at first glance.

ANALYSIS FRAMEWORK:

1. SURFACE INTERPRETATION: What does the riddle seem to be asking about at first glance?

2. MISDIRECTION TRAP: What category of answers would most people naturally think of?

3. COGNITIVE BIAS: What assumptions does the riddle want people to make?

DISTRACTOR CREATION RULES:

Table A1: Per-dimension agreement breakdown on the 5% (22 items) audit set.

| Dimension | $K$ | Units | Disagreements $d_i$ | $D_{o,i} = d_i/22$ | $D_{e,i} = 1 - 1/K$ | $\alpha_i = 1 - D_{o,i}/D_{e,i}$ |
|---|---|---|---|---|---|---|
| Reasoning Type (D1) | 6 | 22 | 3 | 0.1364 | 0.8333 | 0.8364 |
| Trap Type (D2) | 7 | 22 | 2 | 0.0909 | 0.8571 | 0.8939 |
| Cultural Depth (D3) | 2 | 22 | 0 | 0.0000 | 0.5000 | 1.0000 |
| Answer Type (D4) | 11 | 22 | 1 | 0.0455 | 0.9091 | 0.9500 |
| Difficulty (D5) | 3 | 22 | 2 | 0.0909 | 0.6667 | 0.8636 |
| **All** | – | **110** | **8** | **0.07273** | **0.75325** | **0.90345** |

1. Focus on the first main question the riddle seems to ask at first glance.

2. Create answers that fit the OBVIOUS interpretation.

3. Make them plausible for someone who hasn't realized the trick.

4. Include answers from the category people would FIRST think of.

5. Add answers that sound logical but miss the linguistic trick.

6. Avoid answers similar to the correct answer—they must be from different domains.

7. Make someone think "that makes sense" before they realize the trick.

8. The answers must be strictly in বাংলা with no other scripts or languages.

CONSTRAINT REQUIREMENTS (included only when detected by the code):

- CRITICAL: All distractors MUST have exactly {constraints.syllable_count} syllables in Bengali.

- The correct answer "{answer}" has {constraints.correct_syllables} syllables.

- Count carefully: নদী = 2, সাগর = 3, পাহাড় = 3, বাংলাদেশ = 5

- Additional constraints may appear: size, color, shape, time references, etc.

EXAMPLE THINKING PROCESS:

- If the riddle appears to ask about countries, generate country names.

- If it appears to ask about animals, use animal names.

- If it appears to ask about objects, use object names.

REQUIREMENTS:

1. Output distractors strictly in বাংলা with no explanations.

2. Focus on misdirection rather than similarity.

3. Ensure cultural appropriateness for Bengali speakers.

4. Follow detected syllable/letter constraints.

5. Create cognitive traps, not semantic matches.

Output format:

DISTRACTOR_1: বাংলা শব্দ
DISTRACTOR_2: বাংলা শব্দ
…
DISTRACTOR_{n}: বাংলা শব্দ

### A.2.2 Distractor Selection Prompt

You are an expert in cognitive psychology and Bengali riddles. Select the 3 MOST DECEPTIVE distractors that trap people in the riddle's misdirection.

RIDDLE: {question}

CORRECT ANSWER: {answer}

SUGGESTED DISTRACTORS: 1. বিকল্প
2. বিকল্প
3. বিকল্প
…

SELECTION STRATEGY:

Choose distractors that create the strongest cognitive traps, NOT the ones most similar to the correct answer.

EVALUATION CRITERIA:

1. **Misdirection Power**: How well does it exploit the riddle's surface interpretation?

2. **First Instinct Appeal**: Would this be a typical initial guess?

3. **Cognitive Trap Strength**: How convincing is it before someone realizes the trick?

4. **Surface Logic**: Does it make immediate sense?

5. **Diversity**: Prefer distractors from different trap categories.

AVOID:

- Distractors that are too similar to one another.

- Distractors close to the correct answer.

- Obscure or implausible options.

PRIORITIZE:

- Obvious category-based guesses.

- Immediately logical answers.

- Options that delay the “aha!” moment.

Output format:

SELECTED: [comma-separated numbers of the most deceptive options]

## A.3 Zero–Shot Evaluation Prompt and Reproduction Guide

### A.3.1 Full Prompt Template

All models were queried in Bengali with a single-turn user message. For each riddle the script replaces `{question}` and `{options}` with the actual text and candidates (A–D). The prompt is shown below exactly as sent to the API.

PROMPT:

> নিচের ধাঁধাটি সমাধান করুন এবং সঠিক উত্তরের এক অক্ষরে (A, B, C, অথবা D) দিন:
>
> প্রশ্ন: question
>
> বিকল্পসমূহ: A) option_1 B) option_2 C) option_3 D) option_4
>
> শুধু JSON আকারে উত্তর দিন. কোনো ব্যাখ্যা বা বর্ণনা দেবেন না. উদাহরণস্বরূপ: {”উত্তর”: "<আপনার উত্তর এখানে>"}

**English Translation**

> Solve the following riddle and give the correct answer as a single letter (A, B, C, or D):
>
> Question: question
>
> Options: A) option_1 B) option_2 C) option_3 D) option_4
>
> Provide the answer only in JSON format. Do not include any explanation or description. For example: {”Answer”: ”<your answer here>”}

The script enforces a temperature of 0 (except where a provider disallows it) and does not include a system message so that every model receives the same pure zero-shot query.

### A.3.2 Reproducibility Notes

To reproduce the reported zero-shot results:

- **Environment.** Python 3.10+ with the `openai` client library and a valid API key for each provider. Store keys as environment variables (`OPENAI_API_KEY`, `ANTHROPIC_API_KEY`, `NOVITA_API_KEY` etc.).

- **Dataset.** Use the released MCQ JSON file , where each entry contains the riddle, four options, the correct option letter, and the five annotation dimensions.

- **Execution.** Run the provided script and set the provider/model names in the `settings` list. The script automatically handles batching, rate limits, and result logging.

- **Outputs.** For every model a timestamped JSON file is created under `results/zero_shot/`, containing raw model responses, extracted predictions, and per–dimension accuracy statistics.

## A.4 Proprietary Model Evaluation Cost on Zero Shot

Table A2 represents a breakdown of usage cost of proprietary models. Although GPT-5 and Claude-4.1-Opus lead the pack in performance, they come with a significant cost which is several magnitude higher than GPT-4.1 which placed third in overall accuracy. With GPT-5 especially expensive because its lengthy chain-of-thought outputs generate many reasoning tokens that count toward usage fees.

| Model | Cost |
|---|---|
| GPT-4.1 | 0.13$ |
| GPT-5 | 3.40$ |
| Claude-4-Sonnet | 0.43$ |
| Claude-4.1-Opus | 2.19$ |

Table A2: API usage cost of proprietary models

We use a cloud service provider to run evaluation on the open models to reduce infrastructure overhead. However, that is totally optional, as open models are available for free to download and use. As a result, they are not part of API cost analysis.

## A.5 Few–Shot Chain-of-Thought Prompt and Reproduction Guide

### A.5.1 Full Prompt Template

For the hardest but solvable subset of riddles we used a Bengali few–shot chain-of-thought (CoT) prompt that first presents worked examples and then requests a step-by-step analysis before giving the final answer. Below is the exact template; the three Bengali examples remain fixed, while `{question}` and `{options}` are replaced at run time.
PROMPT:
’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’

> আপনি একটি বাংলা ধাঁধার বিশেষজ্ঞ। নিম্ন কিছু ধাঁধার উদাহরণ দেওয়া হলো, যেখানে ধাঁধার সমাধান বিশ্লেষণ করা হয়েছে।
>
> উদাহরণ ১: প্রশ্ন: একটা ঘড়ির উপর দিয়ে একটা ঘোড়া চলে গেল, ঘড়িটার কটা বাজবে। বিকল্পসমূহ: A. সাতটা B. বারোটা C. নটা D. তিনটা যুক্তি: ঘড়ির কাটা ভেঙে যাবে, তাই বারোটা বাজবে। উত্তর: B
>
> উদাহরণ ২: প্রশ্ন: কোন কার চলে না? বিকল্পসমূহ: A. নৌকা B. সাইকেল C. কুকার D. গাড়ি যুক্তি: কুকার যানবাহন নয়, তাই কুকার চলতে পারে না। উত্তর: C
>
> উদাহরণ ৩: প্রশ্ন: নাকের ডগায় পৈতে আটকান চৈতনে মার টান গলায় ধরে দাও পটকান... বিকল্পসমূহ: A. হামানদিস্তা B. লাট্টু C. হাতুড়ি D. দা যুক্তি: বর্ণনা লাট্টুর বৈশিষ্ট্যের সাথে মিলে। উত্তর: B
>
> এখন নিচের ধাঁধাটি সমাধান করুন। প্রথমে যুক্তি ব্যাখ্যা করুন, তারপর উত্তর দিন:
>
> প্রশ্ন: question
>
> বিকল্পসমূহ: options
>
> নিম্নলিখিত ধাপগুলো অনুসরণ করুন: ১. প্রশ্ন বিশ্লেষণ ২. প্রতিটি বিকল্প মূল্যায়ন ৩. যুক্তিসঙ্গত সিদ্ধান্ত
>
> JSON আকারে উত্তর দিন: {”যুক্তি": "<আপনার যুক্তি এখানে>”, ”উত্তর”: ”<A/B/C/D>”}

’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’

Each model received this full text as a single user message, preceded by a system instruction:

> আপনি একটি বাংলা ধাঁধা বিশেষজ্ঞ। সর্বদা ধাপে ধাপে চিন্তা করুন এবং JSON ফরম্যাটে উত্তর দিন।

Temperature was set to 0 when supported.

**English Translated Prompt:**

> You are an expert in Bengali riddles. Below are some example riddles with analyses of their solutions.
>
> Example 1: Question: A horse passes over a clock—what time will the clock show? Options: A. Seven o’clock B. Twelve o’clock C. Nine o’clock D. Three o’clock Reasoning: The clock’s hands will break, so it will show twelve o’clock.[3] Answer: B
>
> Example 2: Question: Which “car” does not move? Options: A. Boat B. Bicycle C. Cooker D. Car Reasoning: A cooker is not a vehicle, so it cannot move. Answer: C
>
> Example 3: Question: “Tie the thread to the tip of the nose, pull it with force, and let it spin around the neck ...” Options: A. Mortar and pestle B. Spinning top C. Hammer D. Machete Reasoning: The description matches the characteristics of a spinning top. Answer: B
>
> Now solve the following riddle. First explain your reasoning, then provide the answer:
>
> Question: question
>
> Options: options
>
> Follow these steps: 1. Analyze the question 2. Evaluate each option 3. Make a logical conclusion
>
> Give the answer in JSON format: {”Reasoning”: ”<Your reasoning here>”, ”Answer”: ”<A/B/C/D>”}

---

[3] Bengali Idiom knowledge is required to understand. In Bengali when it’s 12’0 clock for someone or something that means the person or object is in ruins.

The translated prompt contains few-shot examples with reasoning. However, we must mention that, the translated prompts are given only for transparency. A lot of linguistic and cultural essence of these examples are lost in translation.

### A.5.2 Reproducibility Notes

To reproduce the few-shot CoT results:

- **Environment.** Python 3.10+ with the `openai` client library. Store API keys in environment variables (`OPENAI_API_KEY`, `ANTHROPIC_API_KEY`, `NOVITA_API_KEY` etc.).
- **Dataset.** Use the released the curated JSON subset of riddles identified as “hard but solvable'' based on zero-shot accuracy.
- **Execution.** Run the provided script and edit the `settings` list to specify provider, model name, and rate-limit delays. The script automatically saves JSON results with raw reasoning, extracted answers, and accuracy statistics.
- **Outputs.** Each run produces a timestamped file in `results/chain_of_thought_hard_cases` containing the full model reasoning text and the parsed predictions, enabling direct comparison with the zero-shot evaluation.

## A.6 Trap Type Analysis

**Trap Susceptibility and Reasoning Robustness.** Figure A1 reports model susceptibility to surface-literal misdirection. Although such traps dominate the dataset (358 of 435 riddles), the comparison is still informative for understanding how models handle superficial cues. Claude-Sonnet-4.0 shows the highest vulnerability (47.6%), whereas GPT-5 and Claude-Opus remain lower at 18.7% and 18.4% respectively.

**Performance–Trap Relationship.** As illustrated in Figure A2, overall accuracy and surface-literal susceptibility exhibit a strong negative correlation ($r = -0.89$). Because the surface-literal category is heavily over-represented, this association should be viewed as exploratory rather than conclusive. Nevertheless, the trend hints that models achieving higher accuracy also develop more robust semantic representations that help them resist superficial distractors. We include these results to encourage further, controlled analyses of the relationship between trap type and reasoning depth.

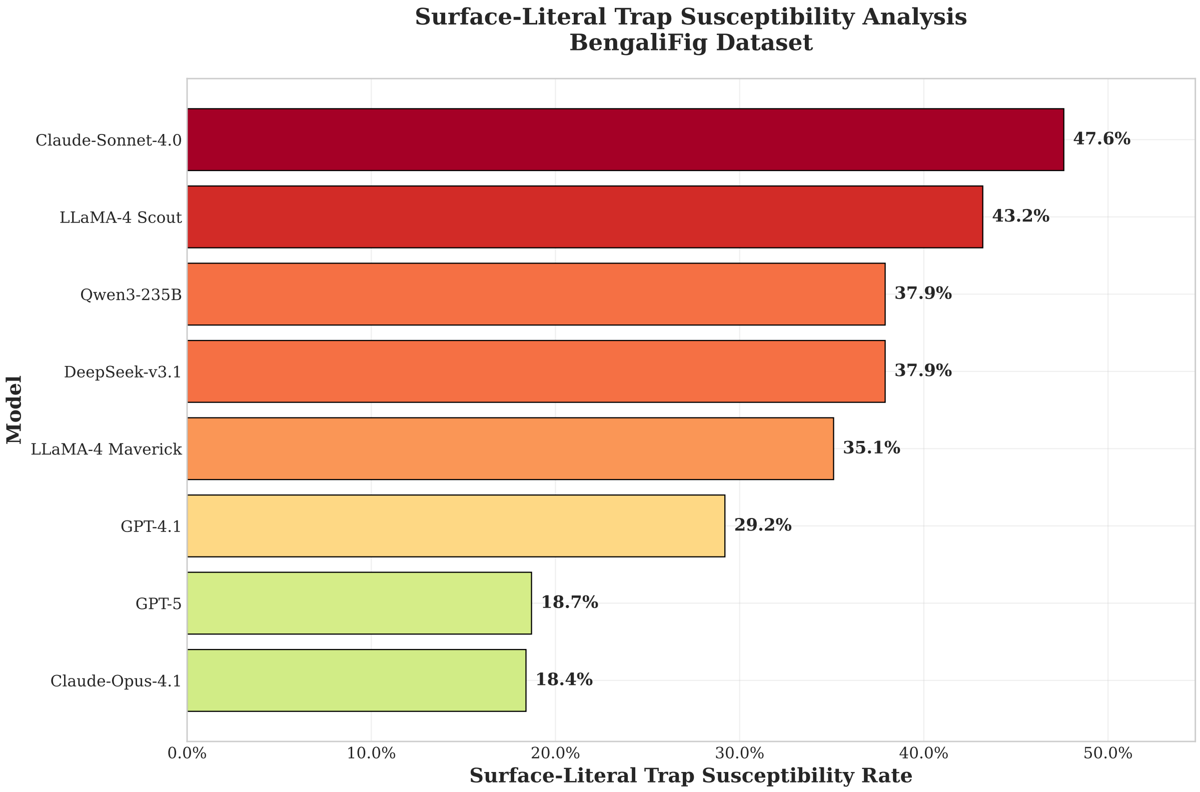


Figure A1: Surface-literal trap susceptibility analysis showing model robustness to misdirection.

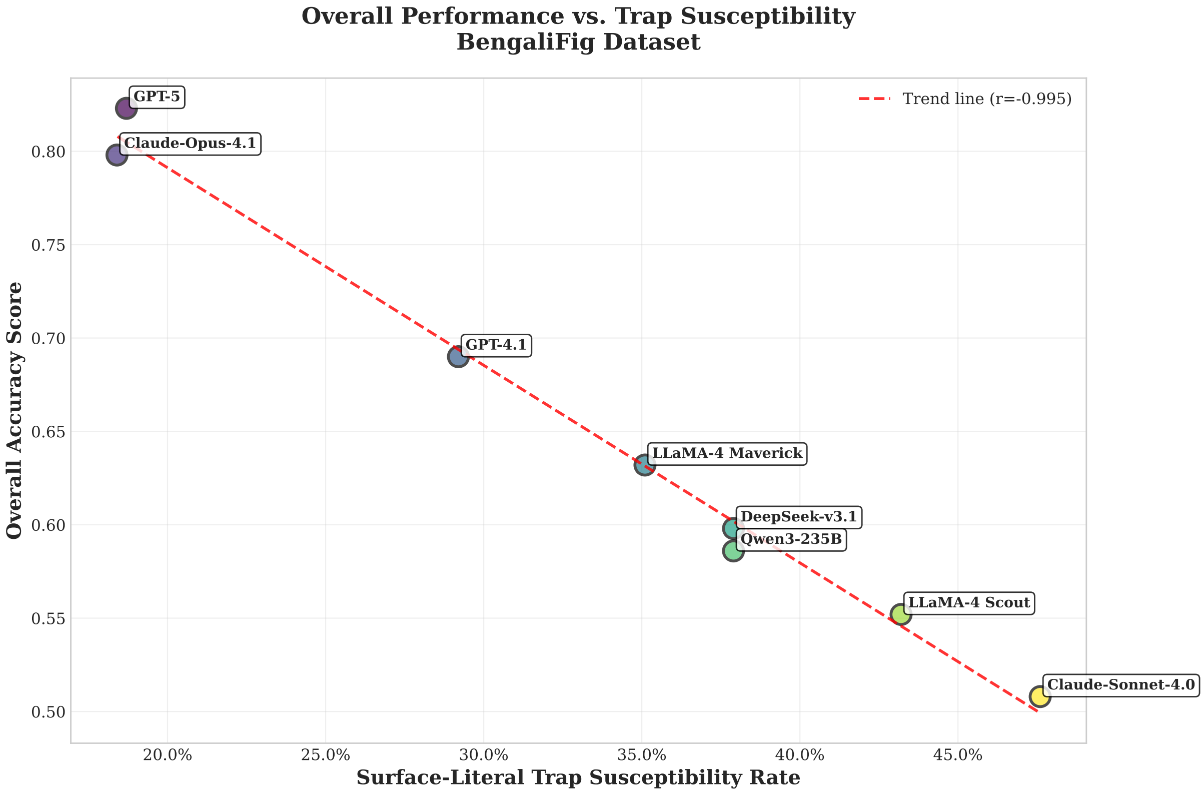


Figure A2: Overall performance versus trap susceptibility, revealing strong negative correlation ($r = -0.89$).

### A.7 LLM Assisted Riddle Annotation

This appendix provides comprehensive details of our LLM-assisted annotation framework used to label the BengaliFig dataset. The annotation process involved two native Bengali speakers who verified and corrected LLM-generated candidate labels across five orthogonal dimensions. We present the complete annotation guidelines provided to human annotators, followed by the exact prompt template used with DeepSeek V3 to generate structured annotation suggestions.

#### A.7.1 Annotation Guidelines

The annotation guidilines provided to the annotators are provided in original text followed by English translation below:

**Original Text:** আপনি একটি বাংলা ধাঁধা টীকাকরণ প্রকল্পে অংশগ্রহণ করছেন। আপনার কাজ হলো প্রদত্ত ধাঁধা-উত্তর জোড়াগুলি বিশ্লেষণ করে পাঁচটি নির্দিষ্ট মাত্রায় উপযুক্ত লেবেল প্রদান করা। প্রতিটি ধাঁধার জন্য একটি এআই সিস্টেম প্রাথমিক টীকা সুপারিশ করবে যা আপনি পর্যালোচনা করে প্রয়োজনে সংশোধন করবেন। অনুগ্রহ করে প্রতিটি ধাঁধা সাবধানে পড়ুন, সাংস্কৃতিক প্রসঙ্গ বিবেচনা করুন এবং আপনার সর্বোত্তম বিচারবুদ্ধি প্রয়োগ করুন।

নিম্নলিখিত নির্দেশাবলী অনুসরণ করে বাংলা ধাঁধাগুলির টীকা প্রদান করুন। প্রতিটি ধাঁধা পাঁচটি মাত্রায় লেবেল করতে হবে:

১. **যুক্তির ধরন (Reasoning Type):**

- **রূপক (metaphorical):** প্রতীকী বা রূপক উপস্থাপনা
- **সাধারণ জ্ঞান (commonsense):** দৈনন্দিন জ্ঞান এবং যৌক্তিক অনুমান
- **বর্ণনামূলক (descriptive):** রূপক ছাড়াই আক্ষরিক বৈশিষ্ট্য
- **শব্দখেলা (wordplay):** ধ্বনি প্যাটার্ন, ভাষাগত বৈশিষ্ট্য
- **যৌক্তিক অনুমান (logical_deduction):** সূত্র থেকে ধাপে ধাপে যুক্তি
- **যৌগিক (compound):** একাধিক যুক্তির ধরন একসাথে

২. **উত্তরের ধরন (Answer Type):**

- **স্থান (place):** অবস্থান (আসাম, ঢাকা, বাংলাদেশ)
- **ব্যক্তি (person):** মানুষ, ভূমিকা (মা, শিক্ষক, রাজা)
- **প্রাণী (animal):** উদ্ভিদ ব্যতীত জীবিত প্রাণী (গরু, পাখি, মাছ)
- **উদ্ভিদ (plant):** উদ্ভিদজগৎ (আম, ধান, ফুল)
- **বস্তু (object):** মানবসৃষ্ট জিনিস (কলম, চেয়ার, বই)
- **প্রাকৃতিক ঘটনা (natural_phenomenon):** প্রকৃতির ঘটনা (বৃষ্টি, রোদ, আগুন)
- **শরীরের অংশ (body_part):** শারীরিক অঙ্গ (চোখ, হাত, মুখ)
- **খাদ্য/পানীয় (food_drink):** ভোজ্য পদার্থ (ভাত, পানি, মিষ্টি)
- **ধারণা (concept):** বিমূর্ত ভাব (ভালোবাসা, সময়, স্বপ্ন)
- **সংখ্যা (quantity):** সংখ্যা, পরিমাপ (তিন, শত, মাইল)
- **টেক্সট/প্রতীক (text_symbol):** লিখিত উপাদান (অ, নাম, চিঠি)

৩. **কঠিনতা (Difficulty):**

- **সহজ (easy):** সরাসরি সংযোগ, সাধারণ জ্ঞান
- **মাঝারি (medium):** মাঝারি চিন্তা, কিছু সাংস্কৃতিক জ্ঞান
- **কঠিন (hard):** জটিল রূপক, গভীর সাংস্কৃতিক অন্তর্দৃষ্টি
- **বিশেষজ্ঞ (expert):** অত্যন্ত বিমূর্ত, বিশেষায়িত জ্ঞান প্রয়োজন

৪. **ফাঁদের ধরন (Trap Type):**

- **আক্ষরিক বিভ্রম (surface_literal):** আক্ষরিক ব্যাখ্যা বিভ্রান্তিকর
- **একাধিক বৈধ (multiple_valid):** বেশ কিছু যুক্তিসঙ্গত উত্তর
- **সাংস্কৃতিক নির্দিষ্ট (culturally_specific):** বাঙালি সাংস্কৃতিক জ্ঞান প্রয়োজন
- **ভাষাগত কৌশল (linguistic_trick):** ধ্বনি/শব্দাংশের প্যাটার্ন গুরুত্বপূর্ণ
- **দিক ভ্রষ্টতা (misdirection):** ভুল ইঙ্গিত
- **প্রাচীন উল্লেখ (archaic_reference):** পুরাতন বাংলা শব্দ/উল্লেখ
- **নেই (none):** কোনো উল্লেখযোগ্য ফাঁদ নেই

৫. **সাংস্কৃতিক গভীরতা (Cultural Depth):**

- **সার্বজনীন (universal):** সাধারণ মানবিক জ্ঞানই যথেষ্ট

- **সাংস্কৃতিক নির্দিষ্ট (cultural_specific):** বাংলা ভাষা জ্ঞান বা সাংস্কৃতিক প্রসঙ্গ অপরিহার্য

**গুরুত্বপূর্ণ বিষয়:**

- বাংলা পাঠ্য সাবধানে পড়ুন, সাংস্কৃতিক প্রসঙ্গ বিবেচনা করুন
- উত্তরের ধরনের জন্য: সবচেয়ে নির্দিষ্ট শ্রেণী ব্যবহার করুন (স্থানের নাম = "স্থান", "বস্তু" নয়)
- সাধারণ বাঙালি বক্তার দৃষ্টিকোণ থেকে কঠিনতা মূল্যায়ন করুন

**English Translation:**

You are participating in a Bengali riddle annotation project. Your task is to analyze the given riddle-answer pairs and provide appropriate labels across five specific dimensions. For each riddle, an AI system will suggest preliminary annotations which you will review and correct as necessary. Please read each riddle carefully, consider the cultural context, and apply your best judgment.

Follow the guidelines below to annotate Bengali riddles. Each riddle must be labeled across five dimensions:

**1. Reasoning Type:**

- **metaphorical:** Symbolic or figurative representation
- **commonsense:** Everyday knowledge and logical inference
- **descriptive:** Literal characteristics without metaphors
- **wordplay:** Sound patterns, linguistic features
- **logical_deduction:** Step-by-step reasoning from clues
- **compound:** Multiple reasoning types combined

**2. Answer Type:**

- **place:** Locations (Assam, Dhaka, Bangladesh)
- **person:** People, roles (mother, teacher, king)
- **animal:** Living creatures except plants (cow, bird, fish)
- **plant:** Vegetation (mango, rice, flower)
- **object:** Man-made items (pen, chair, book)
- **natural_phenomenon:** Nature events (rain, sun, fire)
- **body_part:** Anatomy (eye, hand, mouth)
- **food_drink:** Consumables (rice, water, sweets)
- **concept:** Abstract ideas (love, time, dream)
- **quantity:** Numbers, measurements (three, hundred, mile)
- **text_symbol:** Written elements (letter, name, letter)

**3. Difficulty:**

- **easy:** Straightforward connections, common knowledge
- **medium:** Moderate thinking, some cultural knowledge
- **hard:** Complex metaphors, deeper cultural insight
- **expert:** Highly abstract, specialized knowledge needed

**4. Trap Type:**

- **surface_literal:** Literal interpretation misleads
- **multiple_valid:** Several plausible answers
- **culturally_specific:** Needs Bengali cultural knowledge
- **linguistic_trick:** Sound/syllable patterns matter
- **misdirection:** Red herring clues
- **archaic_reference:** Old Bengali terms/references
- **none:** No significant traps

**5. Cultural Depth:**

- **universal:** General human knowledge sufficient
- **cultural_specific:** Bengali language knowledge or cultural context essential

**Key Points:**

- Read Bengali text carefully, consider cultural context
- For answer type: Use most specific category (place names = ”place”, not ”object”)
- Assess difficulty from typical Bengali speaker perspective

**Illustrative Examples** Representative annotations appear in Table A3. For instance, the riddle "কিভাবে কাঁচা ডিম ফেলে কংক্রিটের ফ্লোর ভাঙা যায়, ডিম না ভেঙে?" is labeled (r,t,a,c,d)=(commonsense, surface-literal, concept, universal, easy).

### A.7.2 LLM Annotation Prompt

The following prompt was used with DeepSeek V3 to generate candidate annotations for Bengali riddles. The LLM receives a riddle-answer pair along with heuristic priors and outputs structured JSON annotations that are subsequently verified by human annotators.

PROMPT:
”””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””

```
You are a Bengali riddle expert.
Annotate this Bengali riddle with 5
labels. Focus on the core cognitive
and cultural aspects.

RIDDLE: তিন অক্ষরের এমন দেশ পেট কাটলে খাই যে
বেশ।
ANSWER: আসাম

ANNOTATION SCHEMA:

1. REASONING_TYPE - Primary thinking
required:

    • metaphorical: Symbolic/figurative
      representation
    • commonsense: Everyday knowledge +
      logical inference
    • descriptive: Literal
      characteristics without metaphors
    • wordplay: Sound patterns,
      linguistic features
    • logical_deduction: Step-by-step
      reasoning from clues
    • compound: Multiple reasoning types
      combined

2. ANSWER_TYPE - What the answer
represents:

    • place: Locations (আসাম,       ঢাকা,
      বাংলাদেশ)
    • person: People, roles (মা,     শিক্ষক,
      রাজা)
    • animal: Living creatures except
      plants (গরু, পাখি, মাছ)
    • plant: Vegetation (আম, ধান, ফুল)
    • object: Man-made items (কলম, চেয়ার,
      বই)
    • natural_phenomenon: Nature events
      (বৃষ্টি, রোদ, আগুন)
    • body_part: Anatomy (চোখ, হাত, মুখ)
    • food_drink: Consumables (ভাত,  পানি,
      মিষ্টি)
    • concept: Abstract ideas (ভালোবাসা,
      সময়, স্বপ্ন)
    • quantity: Numbers, measurements
      (তিন, শত, মাইল)
    • text_symbol: Written elements (অ,
      নাম, চিঠি)

3. DIFFICULTY - Cognitive challenge:

    • easy: Straightforward connections,
      common knowledge
    • medium: Moderate thinking, some
      cultural knowledge
    • hard: Complex metaphors, deeper
      cultural insight
    • expert: Highly abstract,
      specialized knowledge needed

4. TRAP_TYPE - Main misleading element:

    • surface_literal: Literal
      interpretation misleads
    • multiple_valid: Several plausible
      answers
    • culturally_specific: Needs Bengali
      cultural knowledge
    • linguistic_trick: Sound/syllable
      patterns matter
    • misdirection: Red herring clues
    • archaic_reference: Old Bengali
      terms/references
    • none: No significant traps

5. CULTURAL_DEPTH - Cultural knowledge
required:

    • universal: General human knowledge
      sufficient
    • cultural_specific: Bengali
      language knowledge or cultural
      context essential

SUGGESTED VALUES (verify and adjust):
- answer_type: place

KEY POINTS:

    - Read Bengali text carefully,
      consider cultural context
    - For answer_type: Use most specific
      category (place names = "place",
      not "object")
    - Assess difficulty from typical
      Bengali speaker perspective

Output JSON only:
` `json
{
"reasoning_type": "...",
"answer_type": "...",
"difficulty": "...",
"trap_type": "...",
```

| **Riddle (truncated)** | **Answer** | *r* | *a* | *t* | *c* |
|---|---|---|---|---|---|
| কাঁচা ডিম ফেলে কংক্রিট… | কোনভাবেই না | commonsense | concept | surface-literal | universal |
| শুইতে গেলে দিতে হয়… | দরজার খিল | compound | object | culturally-specific | cultural-specific |
| সাহেব কোর্ট প্যান্ট পরে… | পেঁয়াজ | metaphorical | food_drink | surface-literal | cultural-specific |
| ১০ জন মানুষ ১০ ঘন্টায়… | কোন সময়ই না | logical_deduction | concept | surface-literal | universal |
| একটা ঘড়ির উপর দিয়ে… | বারোটা | wordplay | quantity | surface-literal | cultural-specific |

Table A3: Sample riddles with final gold annotations $(r, a, t, c)$.

~~~
"cultural_depth": "...",
"source": "web"
}
` `
~~~

’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’’

We provide here a complete prompt as to how it would be with an example input question-answer pair embedded into it, rather than leaving it as an empty placeholder for transparency and illustrative purpose. The translation and gold annotation after human evaluation is given below:
**Translation Note:** The riddle "তিন অক্ষরের এমন দেশ পেট কাটলে খাই যে বেশ।" translates to ”A three-letter country, when you cut its belly, you eat quite well.” The answer "আসাম" (Assam) is a wordplay where cutting the middle letter "সা" from "আসাম" gives "আম" (mango), which is eaten. Please note that, Bengali graphemes are compound and represents syllables.
**Gold Standard Annotations:** For the example riddle, the LLM suggested the following annotations which were then verified and modified by human annotators:

- `reasoning_type: wordplay`
- `answer_type: place`
- `difficulty: medium`
- `trap_type: surface_literal`
- `cultural_depth: cultural_specific`

# Domain-Specific Adaptation for ASR through Text-Only Fine-Tuning

**Betty Kurian**
Harman International
Bangalore, India
`betty.kurian@harman.com`

**Abhinav Upadhyay**
Harman International
Bangalore, India
`abhinav.upadhyay@harman.com`

**Abhijeet Sengupta**
Harman International
Bangalore, India
`abhijeet.sengupta@harman.com`

## Abstract

Speech recognition models often struggle in specialized domains due to the lack of domain-specific paired audio-text data, making it difficult to adapt general-purpose systems to unique terminology and linguistic patterns. In this work, we propose a text-only domain adaptation method for Whisper, fine-tuning only the decoder using domain-relevant text. Our approach introduces trainable cross-attention bias embeddings, extended with a gated mixture-of-experts routing mechanism, enabling the model to encode domain-specific linguistic priors without any audio data. Unlike ASR adaptation methods that require paired audio-text datasets, our approach is lightweight and resource-efficient. We observe up to a 56% relative improvement in word error rate over the baseline. Our findings demonstrate that text-only adaptation is a practical and effective strategy for improving speech recognition in specialized domains with limited or no domain-specific audio.

## 1 Introduction

Speech recognition technology has advanced significantly in recent years, with applications in virtual assistants, transcription services, and real-time communication systems. These improvements have been driven by supervised learning approaches that rely on paired audio-text datasets to train models capable of mapping language to text Watanabe et al. (2017). Such datasets enable models to learn the complex relationships between speech signals and their textual representations, resulting in robust general-purpose Automatic Speech Recognition (ASR) systems. However, achieving high accuracy in specialized domains remains challenging. Domain-specific ASR systems must address unique linguistic patterns, specialized terminology, and the limited availability of paired audio-text data Bataev et al. (2023). In domains such as healthcare, legal, or scientific research, the limited availability of annotated domain audio constrains the adaptation of general-purpose models, highlighting the need for approaches that reduce reliance on domain-specific audio resources.

To address these challenges, researchers have investigated integrating ASR systems with language models (LMs) through shallow and deep fusion Gulcehre et al. (2015), as well as generating synthetic domain audio using text-to-speech (TTS) systems Huang et al. (2020). Shallow fusion can improve recognition accuracy but requires an external LM during inference, which increases computational cost and latency. Deep fusion incorporates the LM within the ASR training process, but this often demands substantial computational resources and careful tuning to prevent overfitting. TTS-based augmentation provides a way to create domain-specific audio from text, yet the generated speech may contain artifacts and fail to replicate the prosody and acoustic variability of natural speech, limiting its effectiveness for adaptation.

In this work, we propose a text-only domain adaptation method for Whisper, fine-tuning only the decoder using domain-relevant textual corpora. Our approach introduces trainable cross-attention bias embeddings, extended with a gated mixture-of-experts routing mechanism, enabling the model to encode domain-specific linguistic priors without any audio data. This eliminates the dependence on paired domain audio while offering a lightweight and resource-efficient adaptation strategy. We observe up to a 56% relative reduction in word error rate compared to the baseline. These results demonstrate that text-only fine-tuning is a practical and effective approach for improving ASR performance in specialized domains.

The remainder of this paper is organized as follows: Section 2 discusses related work, Section 3 presents the proposed approach, Section 4 details the evaluation and results, and Section 5 concludes

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 78–85
December 23, 2025 ©2025 Association for Computational Linguistics

the paper.

## 2 Related Work

Traditional ASR systems rely heavily on handcrafted features and statistical models, involving multiple stages such as acoustic modeling, phoneme recognition, and language modeling Bell et al. (2020). Recent advances in deep learning, particularly Transformer-based architectures Vaswani (2017), have enabled end-to-end models that map audio directly to text, simplifying the pipeline and achieving state-of-the-art performance. However, adapting these models to new domains remains challenging due to the need for large amounts of labeled audio data, motivating research into more efficient domain adaptation techniques Bell et al. (2020).

A common strategy for domain adaptation is to use text-to-speech (TTS) to synthesize paired speech-text data from target-domain text for fine-tuning ASR models Huang et al. (2020). While effective, this process requires training high-quality multi-speaker TTS models, which is computationally expensive Zheng et al. (2021). To reduce this cost, text-to-spectrogram approaches generate synthetic spectrograms directly from text, removing the need for TTS and audio storage while minimizing the mismatch between synthetic and real audio Bataev et al. (2023). This approach still requires careful training of the spectrogram generator to ensure quality.

Text-only adaptation methods offer a more cost-efficient alternative. These include fine-tuning external language models on target-domain text and integrating them into ASR decoding via shallow fusion Kannan et al. (2018). Contextual biasing methods embed domain-specific phrases to improve recognition of rare terms Aleksic et al. (2015); Chang et al. (2023), while prompt-based techniques condition the ASR model on additional domain cues to guide transcription Suh et al. (2024). Another approach uses pseudo-audio embeddings as prompts for fine-tuning Ma et al. (2024), allowing adaptation without paired data. Tran et al. Tran et al. (2025) propose DAS, a domain adaptation framework that generates domain-specific synthetic speech from LLM-produced text and fine-tunes Whisper with LoRA adapters.

Our approach adapts the ASR model to new domains using text-only fine-tuning, without relying on synthetic audio generation, prompt-based conditioning, or external rescoring. This design reduces computational cost, lowers latency, and simplifies deployment, while enhancing recognition of domain-specific vocabulary. Our method directly adapts the model's language understanding capabilities using only textual data, making it practical for scenarios where domain audio is limited or unavailable.

## 3 Approach

Our approach builds on Whisper, a state-of-the-art ASR model developed by OpenAI Radford et al. (2023). Whisper employs a Transformer-based encoder-decoder architecture, where the encoder processes audio inputs into latent representations, and the decoder generates transcriptions by attending to both the encoder's output and prior textual context. The encoder captures acoustic features such as phonemes, pitch, and rhythm, while the decoder aligns these features with linguistic patterns to produce accurate transcriptions. Whisper's decoder accepts input sequences, enabling the model to incorporate textual descriptions or prompts as part of the input. This feature allows Whisper to condition its transcription generation on additional context, such as domain-specific instructions or metadata. We leverage this capability for domain adaptation by modifying the architecture to focus only on the decoder, bypassing the encoder. This enables adaptation using text-only data without requiring paired audio-text inputs.

The encoder in Whisper generates contextualized representations of the input audio, which are passed to the decoder for processing via the cross-attention mechanism. During cross-attention, the decoder queries the encoder outputs using keys ($K$) and values ($V$), where $K$ represents the contextualized embeddings generated by the encoder, and $V$ serves as the basis for computing attention-weighted outputs that guide the decoder's predictions. We freeze the encoder during training, but the decoder still requires valid K and V representations for the cross-attention mechanism to function correctly, even though the encoder's outputs are no longer updated. To address this, we replace the encoder's output in the cross-attention mechanism with trainable biases $B$. The bias embeddings $B$ denoted as $R^{N \times d}$, where $N$ is the bias sequence length (representing the number of tokens) and $d$ is the embedding dimension, which matches the output dimension of the frozen encoder. These biases

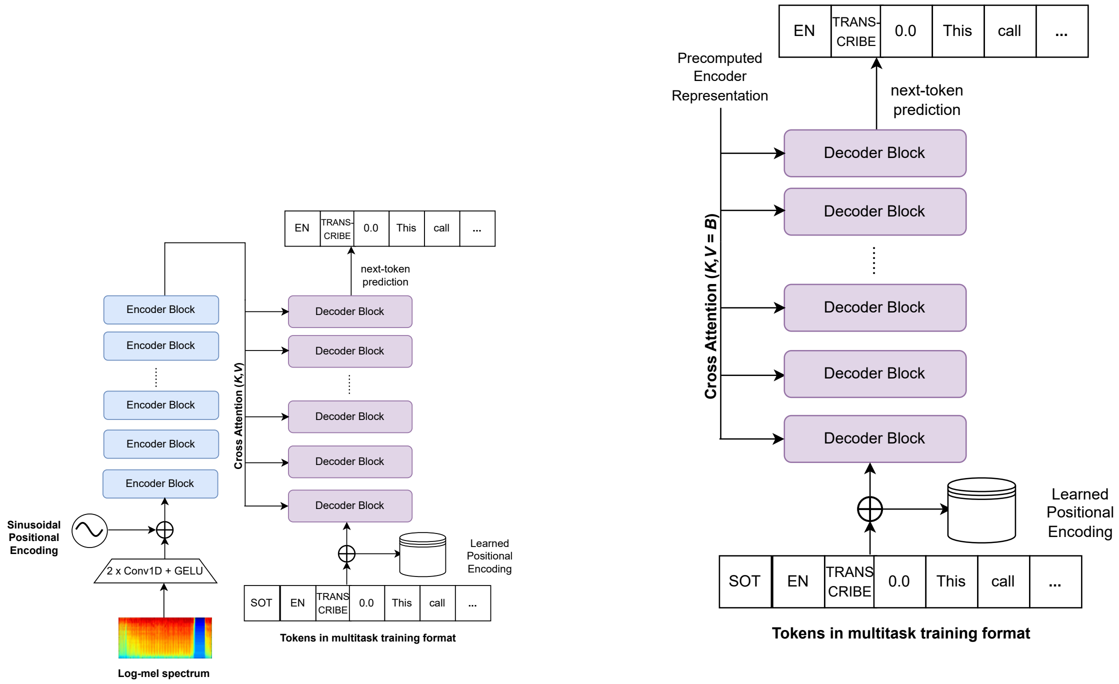


Figure 1: Figure (a) shows the Whisper-base encoder-decoder architecture. Figure (b) shows the modified architecture with domain-specific bias adapters for text-only adaptation (our approach), where multiple expert bias matrices are introduced into the decoder to incorporate domain-specific linguistic priors and guide transcription without using paired audio.

serve as trainable substitutes for the encoder's representations, allowing the decoder to focus entirely on linguistic patterns while maintaining structural compatibility with the original architecture Suh et al. (2024).

We compute cross-attention as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (1)$$

We replace the encoder outputs with the bias embeddings as K and V in the decoder's cross-attention layers. Substituting $K$ and $V$ with $B$, the cross-attention becomes:

$$\text{Cross-Attention} = \text{Attention}(Q, \mathbf{B}, \mathbf{B}) \quad (2)$$

To further refine the contribution of bias embeddings, we introduce a $\tanh$ gating mechanism:

$$G = \tanh(W_g \cdot B) \quad (3)$$

where $W_g$ represents a learnable weight matrix that modulates bias embeddings.

We initialize the bias embeddings $B$ using precomputed representations from the pretrained Whisper encoder.

$$B = E_{pretrained} \quad (4)$$

where $E_{pretrained}$ represents the fixed output of the pretrained Whisper encoder. These bias embeddings are then made trainable during fine-tuning, allowing the model to adapt them for domain-specific text representations. This initialization ensures that the model starts with relevant embeddings while retaining the flexibility to refine them through backpropagation. The trainable biases introduced into the cross-attention layers implicitly capture domain-relevant features during training, allowing the decoder to operate effectively in a text-only setting.

**Gated Routing for Multi-Domain Adaptation:** To handle multiple domain subspaces, we extend the bias embedding design into a *mixture-of-experts* (MoE) framework. Instead of a single bias matrix $B$, we maintain a set of $M$ expert bias matrices $\{\mathbf{B}_m\}_{m=1}^{M}$, each representing domain-specific linguistic priors. A lightweight routing network computes mixture weights $\pi \in R^M$ conditioned on the current decoding context:

$$\pi = \mathrm{softmax}(W_r, \phi(y_{<t})), \tag{5}$$

where $\phi(y_{<t})$ encodes the partial transcription history and $W_r$ is a learnable projection. The aggregated bias is then:

$$\mathbf{B}^* = \sum_{m=1}^{M} \pi_m, \mathbf{B}_m. \tag{6}$$

This $\mathbf{B}^*$ replaces $B$ in the cross-attention mechanism, enabling the decoder to dynamically route attention toward the most relevant domain priors. This structure not only improves adaptation to diverse subdomains but also retains efficiency, as only the small bias matrices and routing parameters are updated during training.

During inference, real audio input is available, and the encoder is reintroduced to generate contextual representations. However, the decoder is trained with bias embeddings, creating a potential mismatch between the learned adaptation and the actual encoder output. To ensure a smooth transition while preserving domain-specific knowledge, we integrate the learned biases with the encoder's output through a linear interpolation.

Given the encoder-generated key and value matrices during inference, we modify the cross-attention mechanism as follows:

$$K' = \alpha K + (1-\alpha)B^*, \quad V' = \alpha V + (1-\alpha)B^* \tag{7}$$

where $K$, $V$ are the encoder's outputs derived from the audio input, $\mathbf{B}^*$ is the aggregated bias embedding from the routing network, and $\alpha$ is a weight that balances the contribution of the encoder and bias embeddings.

The interpolation ensures that the decoder receives both domain-specific cues (from $B$) and actual acoustic representations (from $K$, $V$). We consider the value of $\alpha$ as 0.5.

**Loss Function**: To improve Whisper's performance on domain-specific transcription tasks, we explore alternative loss functions beyond standard cross-entropy. Specifically, we incorporate two loss functions:

1. **Kullback-Leibler (KL) Divergence**: This loss function measures the divergence between two probability distributions, guiding the model towards a better approximation of the true transcription distribution. Minimizing this divergence improves the fluency and accuracy of generated transcriptions.
2. **Bregman Divergence-Inspired Loss**: This loss function prioritizes correct predictions of domain-specific terms (e.g., technical jargon, medical terminology) by assigning higher penalties to errors involving critical domain-specific words.

The combined loss function as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \cdot \mathrm{KL}(P_{\text{true}} || P_{\text{pred}}) + \lambda_{\text{BD}} \cdot \sum_{i=1}^{n} \delta_i \cdot I(w_i \in \mathcal{D}) \tag{8}$$

where $\mathrm{KL}(P_{\text{true}} || P_{\text{pred}})$ is the Kullback-Leibler Divergence between the true transcription distribution $P_{\text{true}}$ and the predicted transcription distribution $P_{\text{pred}}$, $\lambda_{\text{KL}}$ is a hyperparameter that controls the weight of the KL divergence term, $\delta_i$ is a penalty factor for incorrect predictions of domain-specific words, $I(w_i \in \mathcal{D})$ is an indicator function that is 1 if the word $w_i$ belongs to the domain-specific vocabulary $\mathcal{D}$, and 0 otherwise, and $\lambda_{\text{BD}}$ is a hyperparameter controlling the weight of the Bregman Divergence-Inspired loss term. In this work, we construct the domain-specific vocabulary D using public data sources and private data, including domain-specific reports.

## 4 Evaluation and Results

We utilize the Whisper-base English model as the foundation for domain-specific adaptation in our experiments. The domain-specific text data is normalized and tokenized to ensure compatibility with the Whisper tokenizer. We measure performance using Word Error Rate (WER) against the baseline to assess the effectiveness of text-only fine-tuning in improving recognition accuracy in specialized domains.

### 4.1 Dataset

We evaluate the model on three domain-specific datasets:

Earnings Call: The dataset contains quarterly earnings conference calls from S&P 500 companies in 2017 Qin and Yang (2019). It includes domain-specific financial discussions from corporate meetings. For text-only fine-tuning, we use

Table 1: Performance of our model with existing baselines

| Methods | Earnings Call | OCW2 | MedReport |
|---|---|---|---|
| Whisper-base Radford et al. (2022) | 32.9 | 33.5 | 32 |
| Context Pertubation Suh et al. (2024) | 15.15 | **9.79** | NA |
| **Ours** | **14.24** | 17.4 | **18** |

11,736 text files for training and 1,678 audio files for testing, with total audio durations of approximately 30 hours and 5 hours, respectively.

OCW2: The OCW2 dataset from MIT OpenCourseWare covers a range of academic lectures Suh et al. (2024). It contains 24,123 text files for training and 3,447 audio files for testing, corresponding to approximately 40 and 10 hours of audio, respectively.

MedReport: The medical domain often lacks large paired audio-text datasets, but has abundant domain-specific text. We curate a set of medical sentences, including drug and medicine names, from pharmaceutical company annual reports and industry publications. This dataset contains 4,000 text files for training and 1,000 audio files for testing, with total audio durations of approximately 10 and 3 hours, respectively.

### 4.2 Experimental Setup

Training is conducted on an NVIDIA Tesla V100 GPU for a maximum of 1,000 steps, with 200 warmup steps to gradually increase the learning rate. Data loading uses 16 workers, and evaluation is performed every 50 steps. Logging occurs every 10 steps, and model checkpoints are saved every 50 steps. Intermediate evaluations are skipped to accelerate training iterations.

### 4.3 Results and Discussion

We evaluate our approach using Whisper-base as the base model and compare it with two baselines. The first is the pre-trained Whisper-base model Radford et al. (2022). The second is a domain-adapted ASR model trained on paired audio-text data and prompted with LLM-generated descriptions combined with context perturbation Suh et al. (2024).

Model performance is measured using WER. Table 1 presents the overall results, showing that our model achieves notable improvements over both baselines. These gains confirm that text-only fine-tuning enhances recognition accuracy in specialized domains. Table 2 provides example transcriptions for the Earnings Call, OCW2, and MedReport datasets, comparing Whisper-base and our adapted model. After fine-tuning, our model more accurately recognizes domain-specific vocabulary, leading to better transcription quality. In the examples, bold text denotes correct outputs matching the ground truth (previously misclassified by Whisper-base), while underlined text indicates incorrect outputs that differ from the ground truth.

**Earnings Call**: Table 2 compares transcriptions among the ground truth, Whisper-base, and our proposed model (“Ours") for the Earnings Call dataset. In the first example, Whisper-base misinterprets “in fact" as “affect" and “net interest expense" as “that interest expense" while also transcribing “debt" as “dad." The proposed model restores these key financial terms correctly. In the second example, Whisper-base introduces disfluencies such as “you know," misrecognizes “build" as “bill" and distorts “UK" into “u.k.a." In the third example, Whisper-base produces an entirely altered phrase, “many ways when I'm very much focused" deviating significantly from the ground truth, whereas the proposed model correctly outputs “we're now" preserving the intended meaning.

As shown in Table 1, the context perturbation approach improves accuracy, reducing WER to 15.15 on the Earnings Call dataset compared with Whisper-base's 32.9. The proposed method achieves the best performance, lowering WER further to 14.24, corresponding to a 19% relative reduction over Whisper-base and a 0.9% reduction compared with the context perturbation approach.

**OCW2**: Table 2 shows that, for the OCW2 dataset, our model correctly retains technical terms such as “fetch and decode" instead of Whisper-base's “FETCH-ND code" and produces grammatically accurate phrases such as “generator is going to burn out" instead of “generators gonna burn out". For complex scientific content, the proposed model maintains coherence better than Whisper-base, although minor errors persist, such as “phalmus" for “thalamus" (an improvement over Whisper-base's “phalm") and the introduction of “cells that goes" instead of “cells that go".

Table 2: Comparison of the transcription output obtained from Whisper-base and Ours with the ground truth. Bold text represents responses that are correct according to the ground truth, but were misclassified by the Whisper-base model. Incorrect responses are underlined.

| | **Ground Truth** | **Whisper-base** | **Ours** |
|---|---|---|---|
| **Earnings Call** | most of that in fact almost all of that was net interest expense on our automotive debt | most of that affect almost all of that was that interest expense on our automotive dad | most of that, **in fact** almost all of that, was **net** interest expense on our automotive **debt**. |
| | no question about it because customers were trying to decide do they want to build their next datacenter in the uk or should they be building that datacenter someplace else in europe | no question about it because customers were trying to decide that they want to you know bill that next data center in the u.k.a. or should they be building that data center someplace else in europe | no question about it because customers were trying to decide that they want to build that next data center in the **UK** or should they be building that data center someplace else in Europe. |
| | we're now very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses | many ways when I'm very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses. | **we're now** very much focused on operating effectively in a warehouse delivered model which we think we can do because we do it across our other businesses. |
| **OCW2** | generator is going to burn out in let's say 10 or 20 years | generators gonna burn out in let's say 10 or 20 years. | generator **is going to** burn out in let's say 10 or 20 years. |
| | and the fetch and decode stages implement optimizations | and the FETCH-ND code stages implement optimizations | and the FETCH **and Decode** stages implement optimizations |
| | it's the ventral the posterior part of the ventral nucleus the thalamus. and that's where we find the cells that goes to the neocortex as i show there | It's the ventral posterior part of the ventral nucleus of the phalm. And that's where we find the cells that go to the neocortex as I show there. | It's the ventral posterior part of the ventral nucleus of the phalmus. And that's where we find the cells that **goes** to the neocortex as I show there. |
| **MedReport** | Paracetamol is one of the most commonly used medications for pain relief and fever reduction. | Parasetimal is one of the most commonly used medications for pain relief and fever reduction. | **Paracetamol** is one of the most commonly used medications for pain relief and fever reduction. |
| | Azee is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections. | A Z is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections. | **Azee** is a commonly prescribed antibiotic to treat bacterial infections. It contains azithromycin, which is effective against respiratory and skin infections. |
| | Broncol is a bronchodilator that helps manage respiratory conditions like asthma. Cipla's broncol is effective in relieving broncospasm and improving breathing. | Bronkel is a bronco dilator that helps manage respiratory conditions like asthma. Sipla's bronkel is effective in relieving bronchospasm and improving breathing. | Broncal is a **bronchodilator** that helps manage respiratory conditions like asthma. **Cipla's** broncal is effective in relieving broncospasm and improving breathing. |

As reported in Table 1, Whisper-base has a high WER of 33.5. The context perturbation method achieves the lowest WER of 9.79, reflecting strong adaptation when trained with audio data and domain-specific prompts. Our method achieves a WER of 17.4, representing a 16% relative reduction compared to Whisper-base, but not matching the performance of context perturbation due to its reliance on text-only fine-tuning.

The relatively higher WER on the OCW2 dataset arises from the nature of our text-only adaptation strategy. Unlike the context perturbation method, which leverages paired audio–text data and domain audio cues to align acoustic and lexical variations, our approach operates purely on text, without exposure to acoustic or prosodic features present in lecture recordings. OCW2 includes substantial variability in speaker style, pacing, and background conditions, which purely textual fine-tuning cannot capture. Despite this limitation, our model achieves consistent improvements over the base Whisper model, demonstrating that linguistic adaptation alone can transfer domain knowledge effectively even in acoustically complex settings. Future extensions could integrate lightweight audio-conditioned adapters or multi-modal alignment losses to further close this gap.

**MedReport**: For the MedReport dataset, our model shows substantial improvements in recognizing medical terminology. It accurately transcribes “Paracetamol" instead of “Parasetimal" and “Azee" instead of “A Z" which are critical distinctions in medical transcription. Some errors remain, such as “Broncol" being transcribed as “Broncal" but these are less severe than Whisper-base’s phonetic distortions. For example, in a case where “Broncol" appears with additional context about Cipla’s product, the proposed model correctly restores the term. As shown in Table 1, WER improves from 32% to 18% after fine-tuning, corresponding to a 14% relative reduction over Whisper-base.

Across domains, Whisper-base shows frequent structural inconsistencies and misrecognitions that distort meaning. Our model, which relies solely on text-based domain adaptation, produces more accurate and readable domain-specific transcriptions but occasionally hallucinates, especially on short audio segments where insufficient context leads to completions based on statistical likelihood rather than actual input. For example, in OCW2, *“and you can restore this activity. you have a question the intermediate stuff where it’s reduced but not yet denatured how do you”* was transcribed as *“and you can restore this activity. Do you have a question? Yes, so in the intermediate stuff where it’s reduced, but not yet, do you make sure how to use it?”*. Similarly, in Earnings Call, *“the americas were up in midsingle digits with strength in the united states”* became *“America’s Rop in midsingle digits for strengthening audit states.”*. These errors are rare in Earnings Call and mostly substitutions, while OCW2 shows added explanatory phrases. We mitigate hallucinations by appending silence to short segments and applying prompt constraints, improving consistency without requiring audio-text alignment. Overall, the improvements over Whisper-base demonstrate that text-only adaptation can achieve strong domain-specific performance while keeping computational costs low.

## 5 Conclusion

In this work, we present a text-only adaptation method for domain-specific speech recognition by fine-tuning the decoder of the Whisper model. The encoder’s output is replaced with trainable biases, allowing the model to capture domain-specific linguistic patterns without requiring paired audio-text data. The proposed method shows substantial improvements in transcription accuracy, particularly for specialized vocabularies, while maintaining computational efficiency. This demonstrates the practicality of our approach for domain adaptation in settings with limited audio resources. Future work explores integrating fine-tuned small language models (SLMs) with additional modalities, such as video, to further enhance domain-specific recognition performance.

## References

Petar S Aleksic, Mohammadreza Ghodsi, Assaf Hurwitz Michaely, Cyril Allauzen, Keith B Hall, Brian Roark, David Rybach, and Pedro J Moreno. 2015. Bringing contextual information to google speech recognition. In *Interspeech*, pages 468–472.

Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, and Boris Ginsburg. 2023. Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator. *arXiv preprint arXiv:2302.14036*.

Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. 2020. Adaptation algorithms for neural network-based speech recognition: An overview. *IEEE Open Journal of Signal Processing*, 2:33–66.

Shuo-Yiin Chang, Chao Zhang, Tara N Sainath, Bo Li, and Trevor Strohman. 2023. Context-aware end-to-end asr using self-attentive embedding and tensor fusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Yan Huang, Jinyu Li, Lei He, Wenning Wei, William Gale, and Yifan Gong. 2020. Rapid rnn-t adaptation using personalized speech synthesis and neural language generator. In *Interspeech*, pages 1256–1260.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.

Yingyi Ma, Zhe Liu, and Ozlem Kalinli. 2024. Effective text adaptation for llm-based asr through soft prompt fine-tuning. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 64–69. IEEE.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Jiwon Suh, Injae Na, and Woohwan Jung. 2024. Improving domain-specific asr with llm-generated contextual descriptions. *arXiv preprint arXiv:2407.17874*.

Minh Tran, Yutong Pang, Debjyoti Paul, Laxmi Pandey, Kevin Jiang, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, and Xin Lei. 2025. A domain adaptation framework for speech recognition systems with only synthetic data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Xianrui Zheng, Yulan Liu, Deniz Gunceler, and Daniel Willett. 2021. Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5674–5678. IEEE.

# Towards Blind and Low-Vision Accessibility of Lightweight VLMs and Custom LLM-Evals

**Shruti Singh Baghel**[1,*], **Yash Pratap Singh Rathore**[1,*], **Sushovan Jena**[1,*],
**Anurag Pradhan**[2], **Amit Shukla**[1], **Arnav Bhavsar**[1], **Pawan Goyal**[3]

[1]Indian Institute of Technology Mandi
`{s24110, s24036, s20011, amitshukla, arnav}@iitmandi.ac.in`

[2]Vellore Institute of Technology
`anurag.pradhan2023@vitstudent.ac.in`

[3]Indian Institute of Technology Kharagpur
`pawang@cse.iitkgp.ac.in`

[*]These authors contributed equally to this work.

## Abstract

Large Vision-Language Models (VLMs) excel at understanding and generating video descriptions but their high memory, computation, and deployment demands hinder practical use particularly for blind and low-vision (BLV) users who depend on detailed, context-aware descriptions. To study the effect of model size on accessibility-focused description quality, we evaluate SmolVLM2 variants with 500M and 2.2B parameters across two diverse datasets: AVCaps (outdoor), and Charades (indoor). In this work, we introduce two novel evaluation frameworks specifically designed for BLV accessibility assessment: the Multi-Context BLV Framework evaluating spatial orientation, social interaction, action events, and ambience contexts; and the Navigational Assistance Framework focusing on mobility-critical information. Additionally, we conduct a systematic evaluation of four different prompt design strategies and deploy both models on a smartphone, evaluating FP32 and INT8 precision variants to assess real-world performance constraints on resource-limited mobile devices.

## 1 Introduction

Large multimodal vision-language models (VLMs), such as GPT and LLaVA series by OpenAI and Microsoft (Liu et al., 2023; OpenAI, 2023), have shown impressive capabilities in understanding and generating detailed descriptions of visual content. While these large models can produce high quality audio descriptions that align with professional standards, their practical application is restricted by their high computational requirements, dependence on cloud infrastructure, which requires high internet bandwidth making them unsuitable for deployment on everyday devices such as mobile phones or tablets. This renders them impractical for BLV users from experiencing real time, private, on-device accessibility.

Our research investigates if significantly smaller models, which can operate on resource-limited devices, can generate video descriptions that are comparable in quality to those produced by large, resource-heavy ones. In real world settings, BLV users require on device solutions capable of providing timely and detailed descriptions without relying on remote servers or continuous internet connectivity. A lightweight model integrated into a smartphone application could locally process live or prerecorded video, enabling synchronized and context-aware audio feedback such as scene changes, object appearances, and actions delivered directly through headphones.

Small vision-language models are emerging as a promising approach to overcome the drawbacks of larger models while still delivering competitive performance on specific tasks. These compact models, usually with fewer than 2 billion parameters, capable of operating effectively on consumer-grade hardware, enabling on-device implementation and real-time processing. SmolVLM2-500M-Video-Instruct (Allal et al., 2024) and SmolVLM2-2.2B-Video-Instruct (Marafioti et al., 2024) are notable developments in this area, tailored for video understanding tasks.

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 86–94
December 23, 2025 ©2025 Association for Computational Linguistics

Furthermore, human annotations (HA) and contextual information are integrated to enhance model understanding providing comprehensive guidance for accessibility-focused video description generation. However, they frequently fall short for BLV users who need precise, contextually relevant, and in-depth information. To address these limitations, professional audio-description (AD) guidelines developed by organizations such as Netflix, Ofcom, Media Access Canada, and the Described and Captioned Media Program (Li et al., 2025) provide structured frameworks that ensure consistency in character identification, scene description, and narrative flow comprehension. As illustrated in Figure 1, SmolVLM2-500M-Video-Instruct generates increasingly detailed and accessibility-focused descriptions when enhanced with human annotations (HA) and professional AD guidelines.

To validate practical deployment viability, we conducted real-world testing on a mobile device, evaluating both SmolVLM2 variants in FP32 and INT8 precision formats.This on-device deployment approach demonstrates that professional-quality video descriptions can be generated locally on consumer devices without cloud connectivity, establishing feasibility for democratizing video accessibility for BLV users.

**Key Contributions:**

(1)We evaluated SmolVLM2 variants across two different environmental contexts, revealing smaller models often outperform larger variants in specific accessibility scenarios.

(2)We implement four progressive prompting strategies to investigate how instruction complexity affects model performance for BLV users.

(3)We introduce two specialized evaluation frameworks, the Multi-Context Evaluation Framework and Navigation Assistance Framework that address critical gaps in existing evaluation methodologies which currently undervalue BLV users' preferences.

(4)We demonstrate that professional quality audio-descriptions may be produced locally without relying on the cloud through extensive real-world deployment testing on consumer-grade smartphones.

## 2 DATASETS

Our evaluation utilizes two benchmark datasets that represent two different environmental contexts(indoor and outdoor).

- Indoor Dataset (Sigurdsson et al., 2016): From the original 9,848 videos (7,985 training, 1,863 testing), we selected 498 videos and their corresponding human annotations from the test set. This represents approximately 27% of the test set, chosen to include diverse indoor activities while ensuring balanced representation across activity categories (cooking, cleaning activities, etc.).
- Outdoor Dataset (Sudarsanam et al., 2024): We selected 423 outdoor videos and their human annotations from the complete collection of 2,061 clips across all partitions. This 20% sample was stratified across different outdoor scenarios (urban environments, parks, streets, natural settings) to maintain environmental diversity crucial for evaluating outdoor navigation assistance.

  We selected this smaller subset of the dataset to evaluate the model's performance in diverse real-world scenarios with varying lighting, weather, and background complexity.

## 3 Framework

Our research investigates the performance tradeoffs between resource-constrained and resource-intensive vision-language models for accessibility -focused video description. We design a comprehensive evaluation framework that systematically compares SmolVLM2-500M-Video-Instruct (Allal et al., 2024) and SmolVLM2-2.2B (Marafioti et al., 2024) across diverse video content and prompting strategies.

### 3.1 Overview

Our approach enables systematic investigation of how model size affects accessibility-focused video description quality across varying instruction complexity levels. Our experimental design employs four distinct prompting strategies that demonstrate progressive complexity from baseline approaches to comprehensive accessibility-focused instruction integration.

### 3.2 Model Selection

For our evaluation, we selected SmolVLM2-500M-Video-Instruct and SmolVLM2-2.2B-Video-Instruct due to their combined advantages for video description tasks (Marafioti et al., 2024). Both

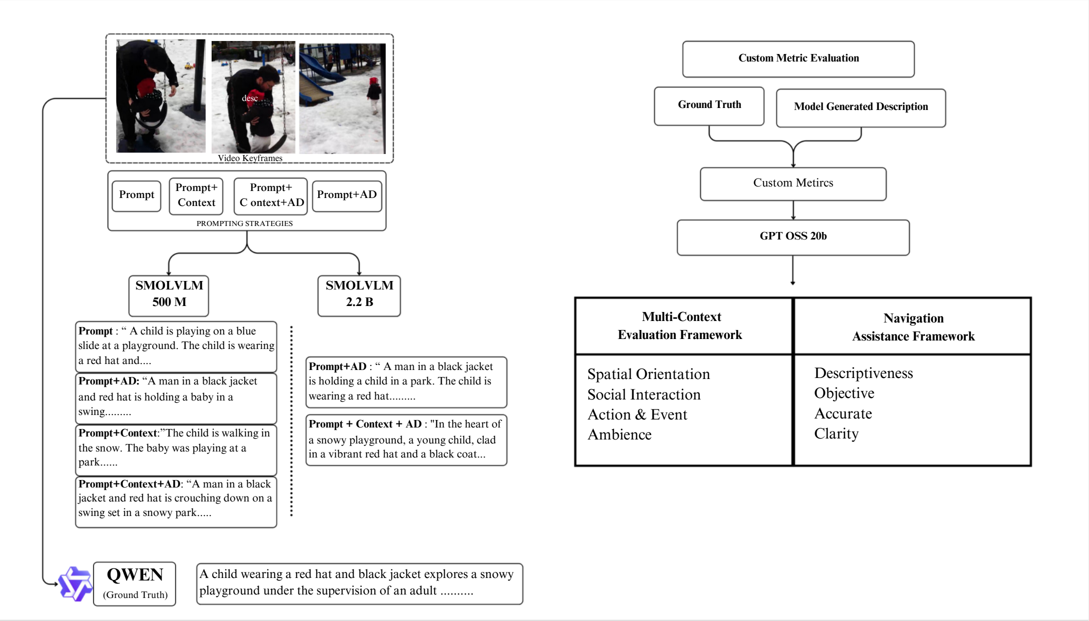

Video Keyframes
desc
Prompt
Prompt+ Context
Prompt+ C ontext+AD
Prompt+AD
PROMPTING STRATEGIES
SMOLVLM 500 M
SMOLVLM 2.2 B
Prompt : " A child is playing on a blue slide at a playground. The child is wearing a red hat and....
Prompt+AD: "A man in a black jacket and red hat is holding a baby in a swing.........
Prompt+Context:"The child is walking in the snow. The baby was playing at a park......
Prompt+Context+AD: "A man in a black jacket and red hat is crouching down on a swing set in a snowy park.....
Prompt+AD : " A man in a black jacket is holding a child in a park. The child is wearing a red hat.........
Prompt + Context + AD : "In the heart of a snowy playground, a young child, clad in a vibrant red hat and a black coat...
QWEN
(Ground Truth)
A child wearing a red hat and black jacket explores a snowy playground under the supervision of an adult ..........
Custom Metric Evaluation
Ground Truth
Model Generated Description
Custom Metircs
GPT OSS 20b
Multi-Context Evaluation Framework
Navigation Assistance Framework
Spatial Orientation
Social Interaction
Action & Event
Ambience
Descriptiveness
Objective
Accurate
Clarity

Figure 1: Experimental Design Overview: Four prompting strategies evaluated across SmolVLM variants and reference model (Qwen). The diagram illustrates progressive complexity from baseline prompt-only approach to comprehensive prompt with context and audio-description guidelines integration. Each strategy generates descriptions that are evaluated against ground truth using both standard NLP metrics and custom accessibility metrics designed for BLV users.

models are explicitly fine-tuned for video understanding with temporal mechanisms essential for coherent description generation, while maintaining exceptional edge deployment viability with GPU memory requirements of only 1.8 GB and 5.2 GB respectively-significantly lower than larger alternatives. The 500M variant achieves competitive performance on Video-MME (42.2) with maximum computational efficiency, while the 2.2B variant offers enhanced quality for scenarios with additional resources, both demonstrating state-of-the-art performance in their respective parameter classes. Critically, both variants support robust instruction following capabilities necessary for implementing professional audio-description guidelines from VideoA11y (Li et al., 2025), enabling real-time inference on consumer hardware and democratizing video accessibility across different computational constraints. Specifically, we address three core research questions:

**Q1.** How effectively can small models match large model performance for accessibility-focused video description when guided by professional audio-description (AD) guidelines?

**Q2.** How do performance trade-offs affect deployment on resource limited hardware such as smartphones?

**Q3.** Why custom accessibility metrics are better than standard NLP metrics in capturing the true preferences of blind and low-vision users for the quality of video descriptions?

This comprehensive evaluation enables us to understand the practical implications of deploying compact VLMs for accessibility applications while maintaining the quality standards necessary for BLV users.

### 3.3 Comprehensive Approach

We use Qwen2.5-VL-7B (Team, 2025) Instruct with expert audio-description guidelines from VideoA11y (Li et al., 2025) to generate ground truth which correctly processes all 42 audio-description guidelines and produces descriptions that meet professional accessibility standards, exhibiting strong instruction following capabilities necessary for putting VideoA11y's methodology into practice.

To efficiently process video content while maintaining essential visual information, we implemented an adaptive keyframe extraction algorithm that analyzes inter-frame differences in the LUV color space. The method computes absolute differences between consecutive frames, applies Hanning window smoothing, and identifies local maxima in the difference signal. In our implementation, we extracted 3-4 keyframes per video, while increasing keyframe density can enhance temporal coverage, it also introduces additional computational cost, a key consideration for on-device deployment.

Following the experimental paradigm established VideoA11y framework (Li et al., 2025), we use four different prompting techniques to assess how contextual information and instruction complexity affect model performance: (1) Prompt Only - utilizing zero-shot generation with the standardized compliant prompt to establish baseline performance without additional guidance. (2) Prompt with Context - incorporating the compliant prompt with original human annotations from the datasets to evaluate the model's ability to leverage existing annotation information. Context refers to ground-truth metadata from the datasets (script-based actions for indoor, audio-visual captions for AV), exactly as implemented in VideoA11y. Human annotations are concatenated with the prompt as "Current Description" before being fed to the MLLM. (3) Prompt with Context and AD Guidelines - combining the prompt with human annotations and 42 professional audio-description guidelines to assess comprehensive multimodal instruction following. (4) Prompt with AD Guidelines - integrating the compliant prompt with audio-description guidelines only to test whether structured accessibility guidelines alone can enable compact models to produce descriptions meeting BLV users' requirements.

### 3.4 Proposed Evaluation Frameworks

Our evaluation protocol addresses the critical limitations of reference based metrics for accessibility applications (Kapur and Kreiss, 2024). We employ dual assessment methodologies: standard NLP metrics for comparison with existing research, and two novel accessibility-centric evaluation frameworks. These frameworks are specifically designed to reflect BLV users' actual needs and preferences. This dual evaluation approach overcomes the systematic bias that reference based metrics exhibit against BLV users' preferences, as demonstrated by Kapur and Kreiss (Kapur and Kreiss, 2024). While VideoA11y effectively assess general description quality, they lack granularity for diverse BLV contexts and navigational needs. To fill these gaps, we introduce two complementary frameworks: Mult-Context BLV Framework and Navigational Assiatance Framework.

#### 3.4.1 Multi-Context BLV Framework

This framework evaluates descriptions across four critical user scenarios that reflect diverse BLV information needs in real-world settings:

(i) Spatial Orientation (1-10 scale): Assesses location descriptions, directional cues, relative positioning, and environmental layout information essential for mental mapping.

(ii) Social Interaction (1-10 scale): Evaluates person identification, interpersonal dynamics, emotional expressions, and social context crucial for understanding human interactions.

(iii) Action & Events (1-10 scale): Measures temporal sequence clarity, activity description completeness, and causal relationships between events.

(iv) Ambience (1-10 scale): Captures mood, lighting conditions, environmental atmosphere, and sensory details that enhance immersive comprehension.

$$\text{MCF\_Score} = \frac{1}{4}\big(S_{\text{spatial}} + S_{\text{social}} + S_{\text{action}} + S_{\text{Ambience}}\big)$$

where:

- $S_{\text{spatial}} \in [1, 10]$: Spatial Orientation Context score
- $S_{\text{social}} \in [1, 10]$: Social Interaction Context score
- $S_{\text{action}} \in [1, 10]$: Action & Event Context score
- $S_{\text{Ambience}} \in [1, 10]$: Ambience Context score

Our framework weights these dimensions based on navigation critical scenarios rather than general description quality.

#### 3.4.2 Navigational Assistance Framework

This framework focuses on mobility critical information through four dimensions essential for spatial navigation and safety:

(i) Descriptiveness: Spatial layout detail, hazard identification, and environmental feature descriptions (obstacles, pathways, boundaries).

(ii) Objectivity: Factual reporting without assumptions, avoiding subjective interpretations of spatial relationships.

(iii) Accuracy: Precision in spatial relationships, object positions, and distance estimations critical for navigation decisions.

(iv) Clarity: Information organization for sequential navigation decision-making, including logical flow and unambiguous directional references.

$$\text{NAF\_Score} = \frac{1}{4}\Big(N_{\text{descriptiveness}} + N_{\text{objectivity}} + N_{\text{accuracy}} + N_{\text{clarity}}\Big)$$

where:

- $N_{\text{descriptiveness}} \in [1, 10]$: Descriptiveness metric score
- $N_{\text{objectivity}} \in [1, 10]$: Objectivity metric score
- $N_{\text{accuracy}} \in [1, 10]$: Accuracy metric score
- $N_{\text{clarity}} \in [1, 10]$: Clarity metric score

#### 3.4.3 Implementation and Validation

For custom accessibility metrics evaluation, we employ GPT-OSS-20B(OpenAI, 2025), a 20-billion parameter open-source language model, following the VideoA11y evaluation methodology (Li et al., 2025). We conducted all evaluations with GPT-OSS-20B running locally via the Ollama(Ollama, 2024) server to ensure offline, reproducible results without network dependencies. The model processes both our Qwen2.5-VL-7B Instruct generated ground truth descriptions and descriptions produced by both SmolVLM variants (500M and 2.2B) using VideoA11y's standardized evaluation template, enabling consistent assessment of the four custom accessibility dimensions. The systematic evaluation enables investigation of our three core research questions outlined earlier (Section 3.2).

### 3.5 Mobile Deployment and Performance Evaluation

To assess real-world deployment viability for accessibility applications, we conducted comprehensive on-device evaluation using a Vivo Y27 smartphone equipped with a MediaTek Helio G85 octa-core processor and Mali-G52 MC2 GPU with 6GB shared system memory. Our deployment methodology employed the llama.cpp framework's llama-mtmd-cli tool, requiring model conversion to .gguf format for mobile compatibility. FP32 variants were converted from their original safetensors format using the official convert_hf_to_gguf.py script, while INT8 quantized versions were generated through Hugging Face's "GGUF My Repo" feature to evaluate precision-performance trade-offs essential for resource-constrained deployment.

The mobile execution environment utilized Termux for Android terminal access, enabling local compilation of llama-mtmd-cli and direct model inference without external dependencies. We implemented a keyframe extraction pipeline using FFmpeg within the mobile environment, processing videos into sequential image frames that were combined with textual prompts incorporating professional AD guidelines.

Both FP32 and INT8 versions of the two models were tested under identical conditions. This setup allowed us to collect detailed performance measurements, including latency, memory usage, and operational behavior during inference on a resource-constrained mobile platform.

## 4 Results and Discussions

All experiments maintain consistent hardware configurations and inference parameters to ensure reproducible comparative analysis between resource constrained and larger models for accessibility focused video description generation.

**Table 1** reveals that SmolVLM2-500M demonstrates strong prompt sensitivity with clear performance patterns across indoor and outdoor scenarios. The Prompt + AD Guidelines approach dominates most evaluation metrics on both datasets, showing consistent alignment with AD-style references and superior lexical overlap performance. However, Prompt + Context + AD Guidelines occasionally excels in semantic-matching metrics like METEOR, indicating that contextual information can enhance meaning preservation. The model shows a notable bias toward AD-style instructions due to reference generation conditions and generally performs better on indoor Charades scenarios compared to outdoor AVCaps environments.

**Table 2** demonstrates that the larger 2.2B model exhibits different contextual utilization patterns. Table 1 presents results for all four prompting strategies using the 500M model

Table 1: SmolVLM2-500M-Video-Instruct: Standard NLP Metrics Performance.

| **Strategy/Dataset** | **BLEU-1** | **BLEU-4** | **METEOR** | **ROUGE-L** | **SPICE** | **CIDER** |
|---|---|---|---|---|---|---|
| *Indoor* | | | | | | |
| Prompt Only | 0.191 | 0.046 | 0.145 | 0.254 | 0.194 | 0.134 |
| Prompt + Context | 0.304 | 0.062 | 0.112 | 0.251 | 0.153 | 0.136 |
| Prompt + AD Guidelines | 0.311 | 0.077 | 0.156 | 0.275 | 0.180 | 0.172 |
| Prompt + Context + AD | 0.287 | 0.070 | 0.153 | 0.268 | 0.173 | 0.194 |
| *Outdoor* | | | | | | |
| Prompt Only | 0.135 | 0.029 | 0.139 | 0.235 | 0.187 | 0.116 |
| Prompt + Context | 0.195 | 0.034 | 0.120 | 0.220 | 0.155 | 0.137 |
| Prompt + AD Guidelines | 0.223 | 0.047 | 0.148 | 0.251 | 0.194 | 0.207 |
| Prompt + Context + AD | 0.273 | 0.055 | 0.162 | 0.247 | 0.171 | 0.131 |

Table 2: SmolVLM2-2.2B-Instruct: Standard NLP Metrics Performance

| **Strategy / Dataset** | **BLEU-1** | **BLEU-4** | **METEOR** | **ROUGE-L** | **CIDEr** | **SPICE** |
|---|---|---|---|---|---|---|
| *Indoor* | | | | | | |
| Prompt + AD Guidelines | 0.2723 | 0.0619 | 0.1353 | 0.2606 | 0.1930 | 0.1768 |
| Prompt + Context + AD Guidelines | 0.3271 | 0.0798 | 0.1363 | 0.2750 | 0.2258 | 0.1841 |
| *Outdoor* | | | | | | |
| Prompt + AD Guidelines | 0.1850 | 0.0345 | 0.1515 | 0.1946 | 0.0884 | 0.1462 |
| Prompt + Context + AD Guidelines | 0.1878 | 0.0331 | 0.1485 | 0.1913 | 0.0719 | 0.142 |

Table 3: SmolVLM2-2.2B-Instruct: Custom Accessibility Metrics Performance

| **Strategy / Dataset** | **Descriptive** | **Objective** | **Accurate** | **Clear** |
|---|---|---|---|---|
| *Indoor* | | | | |
| Prompt + AD Guidelines | 2.508 | 3.25 | 1.935 | 3.345 |
| Prompt + Context + AD Guidelines | 2.529 | 3.246 | 1.78 | 3.414 |
| *Outdoor* | | | | |
| Prompt + AD Guidelines | 2.908 | 2.712 | 1.778 | 3.095 |
| Prompt + Context + AD Guidelines | 2.936 | 2.761 | 1.835 | 3.222 |

whereas Table 2 reports results for the two best-performing strategies:"Prompt + AD Guidelines" and "Prompt+Context+AD Guidelines". This decision was driven by Table 1's clear demonstration that basic "Prompt Only" and "Prompt+Context" strategies consistently underperform compared to AD-enhanced approaches across all standard NLP metrics; therefore we omitted these two strategies for 2.2B model. In indoor scenarios, adding contextual information substantially enhances performance across all metrics, with Prompt + Context + AD Guidelines consistently outperforming the basic AD approach. This indicates the larger model can effectively exploit additional context to improve generation quality in structured, predictable environments. However, in outdoor scenarios, the performance gap narrows significantly, with context sometimes failing to provide meaningful improvements and occasionally diluting performance in precision-focused measures.

**Table 3** examines how contextual integration affects description quality for BLV users. In indoor environments, adding context provides modest improvements in descriptiveness and clarity but introduces slight decreases in objectivity and more notable declines in accuracy, suggesting that enhanced vividness may come at the cost of strict factual reporting. Conversely, in outdoor environments, contextual cues prove particularly valuable, benefiting all evaluation dimensions with especially notable improvements in clarity and accuracy. This pattern indicates that contextual information helps BLV users gain a better awareness of space and is especially helpful in dynamic, visually complex outdoor environments.

**Table 4** reveals distinct strengths between model variants when using optimal prompting strategies. The 2.2B model demonstrates superior clarity and accuracy, along with better objectivity in indoor scenarios, making it more dependable for producing trustworthy, accessible descriptions, while the smaller model excels in descriptive richness.

Table 4: Model Performance Comparison: Prompt + Context + AD Guidelines Strategy using custom metrics

| | **SmolVLM2-500M-Video-Instruct** | | **SmolVLM2-2.2B-Instruct** | |
|---|---|---|---|---|
| **Metric** | **Outdoor** | **Indoor** | **Outdoor** | **Indoor** |
| Descriptive | 3.031 | 2.779 | 2.936 | 2.529 |
| Objective | 2.747 | 2.793 | 2.761 | 3.246 |
| Accurate | 1.719 | 1.627 | 1.835 | 1.780 |
| Clarity | 3.177 | 3.094 | 3.222 | 3.414 |

Table 5: Performance Comparison of SmolVLM2 Models with FP32 and INT8 Quantization

| **Metric** | **SmolVLM2-500M** | | **SmolVLM2-2.2B** | |
|---|---|---|---|---|
| | **FP32** | **INT8** | **FP32** | **INT8** |
| Latency (ms) | 33639.04 | 29904.29 | 2000642.04 | 201306.71 |
| Peak DRAM Usage | 1142.784 MB | 761.856 MB | 2797.216 MB | 2512.896 MB |
| Model Size | 190.22 MB | 103.73 MB | 831.87 MB | 565.05 MB |
| Tokens Per Second *(generation speed)* | 6.41 | 13.55 | 0.05 | 1.47 |
| Time to First Token | 17120.57 ms | 18797.63 ms | 150457.48 ms | 123936.97 ms |
| Time Per Output Token | 155.95 ms | 73.81 ms | 18501.90 ms | 680.44 ms |
| Token Generation Time | 10604.30 ms | 8192.60 ms | 1813186.29 ms | 70085.14 ms |

Table 6: Quantitative Results for Multi-Context Evaluation Framework

| **Model Variant** | **Spatial Orientation** | **Social Interaction** | **Action & Event** | **Ambience** |
|---|---|---|---|---|
| *Outdoor* | | | | |
| 500M | 3.556 | 3.206 | 2.585 | 4.664 |
| 2.2B | 3.416 | 3.271 | 2.632 | 4.925 |
| *Indoor* | | | | |
| 500M | 3.223 | 3.281 | 2.126 | 4.318 |
| 2.2B | 2.976 | 3.332 | 1.949 | 3.532 |

**Table 5** shows results from llama-cpp inference framework. The metrics clearly indicate the low-memory and latency of 500M model over 2.2B. Total latency comprises Load Time (model loading), Prompt Evaluation Time (input processing and tokenization), and Generation Time (step-by-step token generation). Latency depends on both per-token processing speed and the number of generated tokens. For the 500M INT8 model, quantization alters output probabilities due to reduced precision, leading to longer token sequences and increased Generation Time compared to FP32. Although the INT8 model achieves faster per-token processing (73.8 ms/token vs. 155.9 ms/token for FP32), it generates more tokens (111 vs. 68), resulting in higher overall latency.

In **Table 6**, the multi-context evaluation framework shows that model size scaling does not uniformly improve performance across all contextual dimensions for BLV users. The 500M model demonstrates better performance in Ambience context description, showing that the smaller models are good at capturing environmental scenarios and visual mood essential for BLV spatial understanding. The Action & Event context consistently scores lowest across all model-dataset combinations, showing the critical limitation in temporal sequence description that affects BLV users' ability to follow dynamic content.

Three distinct failure patterns are identified via analysis: inaccurate temporal ordering across keyframes, missing causal linkages between events, and collapsing sequential actions into static descriptions ("person near door" instead of "opens, enters, closes door"). Crucially, the Action & Event scores for both model variations are comparable (500M: 2.36 average, 2.2B: 2.29 average), suggesting that our 3 to 4 keyframe sampling rate discussed in Section 3.3 rather than the model's reasoning ability is the bottleneck. Regardless of the number of parameters, this sparse temporal sample only records one frame every two to three seconds, which is

Table 7: Quantitative Results for Navigation Assistance Framework

| Model Variant | Descriptiveness | Objective | Accurate | Clarity |
|---|---|---|---|---|
| *Outdoor* | | | | |
| 500M | 3.570 | 5.107 | 3.147 | 3.930 |
| 2.2B | 3.239 | 4.909 | 3.370 | 3.742 |
| *Indoor* | | | | |
| 500M | 3.258 | 5.02 | 3.002 | 3.533 |
| 2.2B | 2.808 | 5.193 | 3.519 | 3.478 |

insufficient for tracking fast motions.

**Table 7** demonstrate that the 500M model consistently outperforms the 2.2B variant in Objectivity scores, indicating that smaller models provide more factual, assumption-free descriptions crucial for BLV navigation safety. However, the larger model shows better accuracy performance in outdoor scenarios, suggesting enhanced spatial precision in complex environments. The consistently moderate Descriptiveness scores across two models reveal a critical gap in providing the detailed spatial information that BLV users require for effective navigation.

The 500M model excel in BLUE-1, BLEU-4, so the larger model is not strictly better in automatic metrics. It is because the 500M tends to generate shorter, more literal captions which align more closely with the reference wordings and has better n-gram overlap than generated by the larger model. Larger VLMs (2.2B) tend to infer extra context or add own creativity, which LLM-as-judge models may penalize as less "objective" or less "accurate", even if they sound fluent. A 500 M model generates straight and contrained outputs which are easy to parse and thus rated higher on clarity; larger models introduce more complex phrasing that affects the clarity as perceived by the judge (LLM).

The 500M model demonstrates superior performance in outdoor scenarios and achieves higher objectivity scores (5.02-5.11) crucial for BLV safety, while the 2.2B model excels in indoor clarity (3.414 vs 3.094) and spatial accuracy. Action Events score lowest (1.95–2.63) due to VLMs' difficulty with sequential temporal reasoning, while Descriptiveness (2.5–3.6) indicates limited spatial detail for safe BLV navigation.

# 5 CONCLUSION

Our comprehensive evaluation reveals three critical insights that challenge conventional assumptions about model scaling for accessibility applications. We introduce two novel evaluation frameworks "Multi-Context BLV Framework" and "Navigational Assistance Framework" that systematically address the bias of reference-based metrics against BLV user preferences. These frameworks demonstrate that smaller models (500M parameters) often excel in environmental adaptability and objective description generation, while larger models (2.2B parameters) provide enhanced precision in structured scenarios. Mobile evaluation establishes the feasibility of edge deployment with 60 - 83 second inference times for 500M models on consumer hardware, addressing privacy and connectivity barriers that disproportionately affect BLV users. We also discuss the caveats of model scaling and show that smaller models can perform better in metrics than larger when context is focused and resource-limited, not only in latency, also in description quality when the trade-off is not much.

While we demonstrate computational feasibility (latency/memory), our current evaluation does not measure the 'Time-to-Audio' perceived by users, which includes Text-to-Speech (TTS) overhead. Furthermore, our use of GPT-OSS-20B as a judge cannot fully replace human BLV validation. Future work will focus on end-user studies to assess real-world usability. The practical deployment of accessibility-focused VLMs on ubiquitous consumer technology represents a significant step toward democratizing video accessibility, providing BLV users with immediate, private, and contextually relevant video descriptions independent of internet connectivity or centralized services.

# References

Loubna Ben Allal et al. 2024. Smolvlm: A small vision language model. *arXiv preprint arXiv:2504.05299*.

Rhea Kapur and Elisa Kreiss. 2024. Reference-based metrics are biased against blind and low-vision users' image description preferences. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 308–314.

Chaoyu Li, Kohei Watanabe, Elliot Poncin, Nicolas Audebert, Julia Hirschberg, and Heiga Zen. 2025. Videoa11y: Method and dataset for accessible video description. *CHI*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, et al. 2024. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Ollama. 2024. Ollama. Get up and running with large language models locally.

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

OpenAI. 2025. Introducing gpt-oss. Accessed August 2025.

Gunnar A Sigurdsson, Gul Varol, et al. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*.

P. Sudarsanam, I. Martín Morató, A. Hakala, and T. Virtanen. 2024. Avcaps: An audio-visual dataset with modality-specific captions. Zenodo.

Qwen Team. 2025. Qwen2.5-vl.

# Evaluating IndicTrans2 and ByT5 for English–Santali Machine Translation Using the Ol Chiki Script

**Kshetrimayum Boynao Singh**[1,2], **Asif Ekbal**[1], **Partha Pakray**[2]
[1]Indian Institute of Technology Patna, India
[2]National Institute of Technology Silchar, India
`boynfrancis@gmail.com, asif@iitp.ac.in, partha@cse.nits.ac.in`

## Abstract

In this study, we examine and evaluate two multilingual NMT models, IndicTrans2 and ByT5, for English-Santali bidirectional translation using the Ol Chiki script. The models are trained on the MMLoSo Shared Task dataset, supplemented with public English-Santali resources, and evaluated on the AI4Bharat IN22 and Flores test sets, specifically IN22-Gen and Flores200-dev. IndicTrans2 finetune strongly outperforms ByT5 across both directions. On IN22-Gen, it achieves 26.8 BLEU and 53.9 chrF++ for Santali→English and 7.3 BLEU and 40.3 chrF++ for English→Santali, compared to ByT5's 5.6 BLEU and 30.2 chrF++ for Santali→English and 2.9 BLEU and 32.6 chrF++ for English→Santali. On the Flores test set, IndicTrans2 finetune achieves 22 BLEU, 49.2 chrF++, and 4.7 BLEU, 32.7 chrF++. Again, it surpasses ByT5. While ByT5's byte-level modelling is script-agnostic, it struggles with Santali morphology. IndicTrans2 benefits from multilingual pre-training and script unification.

## 1 Introduction

Natural Language Processing (NLP) has made significant progress in a short amount of time, resulting in substantial improvements in machine translation (MT), particularly for languages with extensive resources. However, many low-resource languages (LRLs), particularly those spoken by indigenous and tribal groups, are underrepresented in digital spaces. Santali, a primary tribal language spoken by millions of people in India, Bangladesh, and Nepal, is a good example of this difference. Santali is widely spoken and essential to the culture, but it is still not well represented in digital form, with few language resources available for computational uses, such as machine translation.

A major problem with making strong machine translation systems for Santali is that there aren't many large, high-quality parallel corpora available (lrl, 2025) Conventional machine translation systems, including neural architectures, necessitate extensive bilingual data to proficiently acquire the mapping between source and target languages. The absence of annotated resources significantly hinders the translation quality for Santali, which employs the Ol Chiki script, a writing system that is not extensively supported in conventional NLP tools and tokenisers.

This research addresses these issues by evaluating two sophisticated multilingual translation models, IndicTrans2 (Gala et al., 2023) and ByT5 (Xue et al., 2022), for the bidirectional translation of English–Santali . IndicTrans2 is a transformer-based model that works best with Indian low-resource fo Indic languages (Pakray et al., 2024). It is well known for its ability to transfer information between languages in environments with limited resources. ByT5 works at the byte level, which means it can handle scripts and characters that aren't visible. This suggests that it could enhance language inclusivity and reach a diverse range of linguistic communities.

We use quantitative metrics, such as BLEU and chrF++, to evaluate the quality of translations into the Ol Chiki script. Our research aims to demonstrate the versatility of models for underrepresented languages and to promote the development of more inclusive and accessible machine translation systems for low-resource linguistic communities.

## 2 Related Works

Research in machine translation (MT) for low-resource (Singh et al., 2023b) languages (LRLs) has gained significant momentum in recent years, as the NLP community increasingly focuses on linguistic inclusivity and equitable access to technology. Early approaches to MT primarily relied on rule-based and statistical methods, which required extensive linguistic expertise and manually crafted translation rules. Although these systems were innovative for their time, they often suffered from scalability issues and produced suboptimal results for languages with limited parallel corpora (Singh et al., 2024b) and sparse digital resources.

The emergence of neural machine translation (NMT) (Appicharla et al., 2024) has revolutionised

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 95–100
December 23, 2025 ©2025 Association for Computational Linguistics

the field by introducing deep learning architectures capable of modelling complex language patterns. The Transformer model and its subsequent variants demonstrated remarkable improvements in translation fluency and accuracy, particularly for high-resource languages. However, NMT models remained heavily data-dependent, and their effectiveness diminished substantially for low-resource languages that lacked sufficient training data.

To address this limitation, researchers have begun developing multilingual NMT models (Singh et al., 2024a) capable of learning shared representations across multiple languages. Such models leverage (Singh et al., 2023a) cross-lingual transfer (Wei et al., 2024) learning, enabling knowledge gained from high-resource languages to improve translation quality for related low-resource ones. IndicTrans2 represents one such advancement tailored explicitly for Indian languages (Dabre and Kunchukuttan, 2024). It employs a shared multilingual (Limisiewicz et al., 2024) encoder–decoder architecture, effectively utilising linguistic similarities among Indo-Aryan and Dravidian languages to enhance performance for low-resource pairs.

In contrast, ByT5 belongs to a newer generation of models that operate at the byte level, eliminating the need for language-specific tokenisation. By processing text as raw bytes, ByT5 can seamlessly handle diverse writing systems and scripts, including those that are poorly represented in mainstream tokenisers. This makes it particularly advantageous for languages like Santali, which is written in the Ol Chiki script, a script not widely supported in traditional NLP pipelines.

Our work builds upon these foundational advances by applying and comparing IndicTrans2 and ByT5 to the English–Santali translation task. Through this comparative analysis, we aim to evaluate how multilingual and byte-level (Nehrdich et al., 2024) modelling approaches perform on a genuinely low-resource (Bhaskar and Krishnamurthy, 2024) language with unique orthographic and linguistic characteristics.

## 3 Linguistics of Santali

Santali[1] is one of the Munda languages in the Austroasiatic language family. Other Munda languages include Mundari, Ho, and Korku. The Santal community, which is one of the largest indigenous tribal groups in India, speaks it most of the time. The Santali-speaking population is estimated to be around 7 million, primarily residing in the Indian states of Jharkhand, West Bengal, Odisha, and Bihar. There are also smaller groups of speakers in Assam and neighbouring countries, such as Bangladesh and Nepal.

In the past, Santali was written in several scripts, including Devanagari, Bengali, Odia, and Latin-based orthographies. This was because of regional language influences and colonial legacies. In 1980, however, the Ol Chiki script, which Pandit Raghunath Murmu developed in the mid-20th century, was officially recognised as the standard writing system for Santali. Ol Chiki was made to better show the phonological structure of the language than borrowed scripts, which had problems with sound representation and spelling consistency.

The use of Ol Chiki has been very helpful in reviving Santali, maintaining its vitality, and making it more consistent. It has helped the language get more attention in formal education, literature, and online spaces. Alternative scripts are still used informally, particularly in regions with multiple languages. However, the Government of India now officially recognises Ol Chiki and supports it in the Unicode Standard, ensuring it works with modern computer systems.

To sum up, Santali is a vibrant language with numerous forms and rich cultural significance for the Santal people. The institutionalisation of Ol Chiki is a crucial step toward preserving its linguistic identity and ensuring it is represented in the digital and technological age.

## 4 Methodology

In this study, we delineate the methodology in two principal phases: dataset preprocessing and model architecture. First, we explain how we made and organized the English-Santali parallel dataset for training. Second, we examine the architecture of the two models used, IndicTrans2 and ByT5, and explain how each is utilised to translate from English–Santali.

### 4.1 Datasets

The dataset utilised in this study significantly expands upon the original 20,000 English–Santali parallel sentences provided by the MMLoSo 2025

---

[1] `https://en.wikipedia.org/wiki/Santali_Wikipedia`

| Split | Sentence | English | Santali |
|---|---|---|---|
| Train Set | 104,451 | 1,082,726 | 1,036,528 |
| Valid | 1503 | 14,850 | 16,001 |
| Test IN22 | 1024 | 25,348 | 26,676 |
| Test Flores | 997 | 20,955 | 22,912 |
| **Total** | **107,975** | **1,143,879** | **1,102,117** |

Table 1: Dataset statistics for the English–Santali corpus. The table reports the number of sentences and total tokens in English and Santali across the training set, IN22-Conv validation set, IN22-Gen test set, and Flores200-dev test set.

| Category | Details for ByT5 Model finetune |
|---|---|
| **ByT5 Training Configuration** | |
| Model | `google/byt5-small` (byte-level) |
| Batch Size | 64 |
| Learning Rate | $3 \times 10^{-4}$ |
| Epochs | 5 |
| Max Length | 256 characters (source & target) |
| **Evaluation Metrics** | |
| Metrics | BLEU, chrF++, CER (Character Error Rate) |
| **Important Flags** | |
| `data_dir` | Path to dataset (required) |
| `output_dir` | Directory to save model (required) |
| `fp16` | Enable mixed precision training |
| `lowercase` | Normalize all text to lowercase |

Table 2: Training configuration for ByT5 finetuning.

| Category | Details for IndicTrans2 finetune |
|---|---|
| **Optimization** | |
| Optimizer | Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$) |
| Learning Rate | $3 \times 10^{-5}$ with inverse square root decay |
| Warmup | 2,000 steps (from $1 \times 10^{-7}$) |
| Gradient Clipping | 1.0 |
| **Training Configuration** | |
| Batch Size | 32,768 tokens (effective across 8 GPUs) |
| Max Updates | 100,000 |
| Mixed Precision | FP16 |
| **Regularization** | |
| Dropout | 0.2 |
| Label Smoothing | 0.1 |
| Early Stopping | Patience of 10 checkpoints (BLEU-based) |

Table 3: Training configuration for IndicTrans2 finetuning.

organisers [2]. To enhance model robustness and improve translation quality, we incorporate publicly available English–Santali resources[3], followed by data augmentation techniques to increase corpus diversity further. After integration and preprocessing, the final dataset, as presented in Table 1, comprises 107,975 parallel sentences, covering a broad spectrum of domains, including culture, daily communication, news-style texts, and general knowledge.

All sentence pairs were manually inspected for alignment quality. Comprehensive preprocessing was applied, including text cleaning, Unicode normalisation for the Ol Chiki script, tokenisation, and consistency checks to ensure high-quality parallelism. The expanded corpus reflects Santali's rich morphology, with the Santali side containing 1,102,117 tokens, compared to 1,143,879 tokens on the English side.

The dataset is divided into a training set, IN22-Conv validation set, and two test sets (IN22-Gen and Flores200-dev). Table 1 provides the complete dataset breakdown. On average, English sentences contain 14–16 tokens, whereas Santali sentences contain 17–19 tokens, indicating Santali's morphologically richer structure. The dataset and related resources are available.[4]

### 4.2 Model Architecture

**ByT5 Architecture:** The proposed English-Santhali translation system is based on the ByT5 architecture Table 2, a token-free version of the T5 and mT5 (Xue et al., 2021) transformer models. It was designed to work directly on raw UTF-8 bytes, rather than using subword tokenisation. This design doesn't use word tokens, allowing the model to handle any language script, including low-resource and morphologically complex languages such as Santhali, which can be written in multiple scripts, like Ol Chiki and Devanagari. ByT5 has a small vocabulary of only 256 bytes. It uses a sequence of byte values to represent each character, which solves problems with words that aren't in the language, spelling mistakes, and noisy input. The model is based on an encoder-decoder framework, but it uses an unbalanced "heavy encoder" architecture, where the encoder is three times deeper than the decoder. The model can learn a soft lexicon by understanding word structure, morphology, and syntactic patterns directly from byte sequences thanks to this deeper encoder. The decoder, on the other hand, focuses on making coherent target text. Additionally, ByT5 relocates the parameters that traditional models utilise for large token embed-

---

[2] https://www.kaggle.com/competitions/mmloso2025/data?select=santali-train.csv

[3] https://opus.nlpl.eu/results/en&sat/corpus-result-table

[4] https://github.com/helloboyn/MMLoSo25-IT2-BT5-ES-MT

| **Testset** | **Model** | **BLEU (En→Sa)** | **chrF++ (En→Sa)** | **BLEU (Sa→En)** | **chrF++ (Sa→En)** |
|---|---|---|---|---|---|
| **IN22-Gen** | IndicTrans2-baseline | 5.5 | 35.8 | 24.8 | 51.0 |
| | IndicTrans2-finetuned | **7.3** | **40.3** | **26.8** | **53.9** |
| | ByT5 Model | 2.9 | 32.6 | 5.6 | 30.2 |
| **Flores200-dev** | IndicTrans2-baseline | 3.3 | 29.5 | 19.5 | 45.1 |
| | IndicTrans2-finetuned | **4.7** | **32.7** | **22.0** | **49.2** |
| | ByT5-finetune | 2.7 | 23.7 | 6.1 | 26.7 |

Table 4: BLEU and chrF++ evaluation scores for IndicTrans2 (baseline and finetuned) and ByT5 on the IN22-Gen and Flores200-dev test sets for both translation directions.

dings to its transformer layers, making them more powerful and efficient. The model is pre-trained with a span corruption objective that hides longer byte spans to help it understand the context better. When fine-tuning for English to Santhali translation, the encoder takes English sentences and turns them into byte sequences. The decoder then makes the Santhali translation one byte at a time. The ByT5 architecture offers several benefits, including the ability to work with any script, handle noisy or unseen inputs, and generalise effectively. This makes it a great choice for building reliable translation systems for low-resource languages, such as Santhali.

**IndicTrans2 Architecture:** IndicTrans2 (Gala et al., 2023) is an innovative multilingual neural machine translation (NMT) model that uses the Transformer architecture 3. A shared multilingual encoder-decoder framework enables it to work with a wide range of Indian languages. The model improves translations by leveraging the fact that Indian languages share similarities with each other. This is especially true for pairs with limited resources, such as English–Santali. The main things about IndicTrans2 are: The Transformer model is the basis for IndicTrans2. It utilises self-attention mechanisms to identify both short- and long-range dependencies within sentences. This enables the model to produce translations that are both fluent and contextually relevant. Multilingual Training: The model learns from a large amount of data that is similar across different Indian languages. This lets it learn shared representations, which helps it generalise better and makes it easier for high-resource languages to share information with low-resource languages. Shared Embeddings: IndicTrans2 utilises shared embedding spaces across languages to leverage patterns that are common to both languages. This method improves the ability to translate between languages that are significantly different from each other. Fine-Tuning Capability: The model can be fine-tuned for specific language pairs, allowing it to adjust to the unique syntax, morphology, and script of languages with limited resources, such as Santali.

## 5 Evaluation Metrics and Analysis

We evaluate our English to Santali and Santali to English translation models using two standard automatic metrics: BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017). BLEU measures n-gram precision and is widely used for machine translation quality estimation, whereas chrF++ captures character-level similarity and is particularly effective for morphologically rich languages such as Santali.

Table 4 reports the results on two benchmark datasets, IN22-Gen with 1,024 sentences and Flores200-dev with 997 sentences. The comparison includes three systems: IndicTrans2-baseline, IndicTrans2-finetuned, and ByT5.

Across both datasets and translation directions, IndicTrans2-finetuned consistently achieves the best performance. On the IN22-Gen set, the model attains 7.3 BLEU and 40.3 chrF++ for English to Santali, and 26.8 BLEU and 53.9 chrF++ for Santali to English. In contrast, ByT5 yields 2.9 BLEU and 32.6 chrF++ for English to Santali, and 5.6 BLEU and 30.2 chrF++ for Santali to English. These results indicate that byte-level modelling is less effective in handling the Ol Chiki script and Santali morphology.

The same pattern is observed on Flores200-dev. IndicTrans2-finetuned achieves 4.7 BLEU and 32.7 chrF++ for English to Santali, and 22.0 BLEU and 49.2 chrF++ for Santali to English. ByT5 again performs considerably lower, reinforcing that subword-based multilingual pretraining is more suitable for this low-resource language pair.

Overall, the findings clearly demonstrate that IndicTrans2, especially when finetuned, provides superior translation quality for English–Santali, offering stronger lexical accuracy, improved handling of the Ol Chiki script, and better character-level consistency compared to ByT5.

## 6 Conclusion

In this work, we investigated English–Santali machine translation using the Ol Chiki script and conducted a focused comparison between two multilingual models: ByT5 and IndicTrans2. By isolating the English–Santali pair, we provided a clear assessment of model performance on this low-resource language. Our results show that the finetuned IndicTrans2 model consistently delivers higher BLEU and chrF++ scores than ByT5 across multiple benchmarks and translation directions. These findings highlight the advantages of subword-based multilingual pretraining for handling Santali's morphology and script-specific characteristics. Overall, our study demonstrates that careful model selection and targeted finetuning play a crucial role in improving translation quality for low-resource languages, contributing to broader efforts toward digital inclusion and linguistic preservation.

## Acknowledgement

The authors express gratitude to the COIL-D Project under Bhashini, funded by MeitY, for the support and resources that facilitated the successful execution of this research.

## References

2025. Multimodal models for low-resource contexts and social impact 2025. Kaggle Competition.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Yash Bhaskar and Parameswari Krishnamurthy. 2024. Yes-MT‘s submission to the low-resource Indic language translation shared task in WMT 2024. In *Proceedings of the Ninth Conference on Machine Translation*, pages 788–792, Miami, Florida, USA. Association for Computational Linguistics.

Raj Dabre and Anoop Kunchukuttan. 2024. Findings of WMT 2024‘s MultiIndic22MT shared task for machine translation of 22 Indian languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 669–676, Miami, Florida, USA. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Nehrdich, Oliver Hellwig, and Kurt Keutzer. 2024. One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13742–13751, Miami, Florida, USA. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024a. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for

English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024b. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

Bin Wei, Zheng Jiawei, Zongyao Li, Zhanglin Wu, Jiaxin Guo, Daimeng Wei, Zhiqiang Rao, Shaojun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie, and Hao Yang. 2024. Machine translation advancements of low-resource Indian languages by transfer learning. In *Proceedings of the Ninth Conference on Machine Translation*, pages 775–780, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

# Challenge Track: LoRAs in All Directions - Directional Adapters and Noisy-Channel Reranking for Indic MT

**Sajay Raj**
Woxsen University
`sajayraj08@gmail.com`

## Abstract

Low-resource machine translation for Indic languages remains challenging, especially when high-resource languages such as Hindi and English must be translated to and from very low-resource, grammatically rich languages like Bhili, Mundari, Santali, and Gondi.

We describe our winning system for a recent shared task in this setting. We start from a strong pretrained Indic MT backbone, IndicTrans2, and fine-tune it jointly on all translation directions, pushing the model close to memorization under strict data constraints. On top of this backbone, we add direction-specific low-rank adapters based on LoRA that allow each language pair to specialize while still sharing most parameters. At inference time, we further couple these directional adapters through a noisy-channel objective, in which forward and reverse models jointly score a set of candidate translations, encouraging outputs that are both fluent in the target language and informative about the source.

This combination of shared pretraining, directional parameter-efficient adaptation, and noisy-channel reranking substantially improves over a strong fine-tuned baseline. We release our codebase at `https://github.com/SajayR/LoRA-in-All-Directions`.

## 1 Introduction

The MMLoSo shared task (mml, 2025) targets a difficult gap in machine translation: bridging high-resource languages (Hindi, English) with extremely low-resource community languages (Bhili, Mundari, Santali, Gondi). These languages are particularly challenging for standard models because they utilize diverse scripts (such as Ol Chiki) and complex word structures, yet lack the large-scale parallel data required to learn these features effectively.

In this regime, standard training strategies face a dilemma. Training separate models for each direction creates data fragmentation, leading to poor convergence. Conversely, joint multilingual fine-tuning maximizes transfer learning but introduces *interference* (or "negative transfer"), where the model's capacity is dominated by high-resource directions, often resulting in script hallucinations or morphological simplification in the lower-resource targets.

Our winning submission addresses this trade-off through a *saturate-then-specialize* strategy. We hypothesize that while a shared backbone is necessary to learn general linguistic representations, distinct parameter spaces are required to resolve orthographic and grammatical conflicts. We therefore combine massive joint fine-tuning (to saturate the backbone with domain knowledge) with direction-specific Low-Rank Adapters (LoRA) (to isolate task-specific constraints) (Hu et al., 2022). Finally, to counter the semantic drift common in low-resource generation, we abandon greedy decoding in favor of a noisy-channel formulation (Pang et al., 2022), using the LoRA adapters to strictly enforce mutual consistency between source and translation.

## 2 Task and Data

### 2.1 Shared task setup

The shared task focuses on translation between two high-resource languages (Hindi, English) and four low-resource Indic languages: Bhili, Mundari, Santali, and Gondi. Let

$$\mathcal{L} = \{\text{Hindi}, \text{English}, \text{Bhili}, \\ \text{Mundari}, \text{Santali}, \text{Gondi}\}$$

be the set of languages.

A translation direction is defined as an ordered pair $d = (\ell_s \rightarrow \ell_t)$ with $\ell_s, \ell_t \in \mathcal{L}$. The task provides parallel datasets $\mathcal{D}_d = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_d}$ for each direction, where source $x$ and target $y$ are in their native scripts. The objective is to produce a translation $\hat{y}$ given $x$ and the direction $d$.

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 101–105
December 23, 2025 ©2025 Association for Computational Linguistics

### 2.2 Leaderboard Score

The leaderboard reports a single scalar score $S$ that mixes BLEU(Post, 2018) and chrF(Popović, 2015) across all eight directions.

Let $\mathrm{BLEU}_{H\to L}$ be the mean BLEU over the four high→low directions (Hin→Bhi, Hin→Mun, Hin→Gon, Eng→San), and let $\mathrm{BLEU}_{L\to H}$ be the mean over the four low→high directions (Bhi→Hin, Mun→Hin, Gon→Hin, San→Eng). We define $\mathrm{chrF}_{H\to L}$ and $\mathrm{chrF}_{L\to H}$ analogously, replacing BLEU with chrF.

The final score is

$$S = 0.6\big(0.6\,\mathrm{BLEU}_{H\to L} + 0.4\,\mathrm{BLEU}_{L\to H}\big) \quad (1)$$
$$+\,0.4\big(0.6\,\mathrm{chrF}_{H\to L} + 0.4\,\mathrm{chrF}_{L\to H}\big). \quad (2)$$

BLEU contributes 60% of $S$ and chrF 40%; within each metric, high→low directions get 60% of the weight and low→high directions 40%.

Two variations of this score are reported: the **Public Score**, calculated on a fixed 25% subset of the test data during the competition, and the **Private Score**, calculated on the remaining 75% hidden subset to determine the final rankings.

## 3 Methodology

## 4 Backbone Selection

We initially compared **NLLB-200** (600M) (NLLB Team et al., 2022) and **IndicTrans2** (360M) (Gala et al., 2023) by fine-tuning both for 100k steps. Table 1 shows that IndicTrans2 outperformed NLLB by nearly 9 points despite being half the size.

| Backbone | Public Score |
|---|---|
| NLLB-600M | 243.18 |
| IndicTrans2-360M | **252.11** |

Table 1: Leaderboard scores at 100k steps.

Tokenization analysis (Table 2) reveals that NLLB suffers from higher word fragmentation across the board. In contrast, IndicTrans2 offers far superior tokenization stability. While this comes at the cost of a slightly higher unknown token rate (Table 2), the difference is negligible ($<1.7\%$ worst-case) and easily mitigated via decoding constraints. We therefore proceed with IndicTrans2.

### 4.1 Tag-Only Preprocessing

To minimize pipeline complexity and avoid brittle external preprocessors for these under-resourced languages, we adopt a "tags-only" preprocessing strategy. We avoid script unification or transliteration. Instead, we condition the model purely via tag prefixing (Johnson et al., 2017).

| | Fertility (↓) | | Unknown tokens % (↓) | |
|---|---|---|---|---|
| **Language** | **NLLB** | **IndicTrans2** | **NLLB** | **IndicTrans2** |
| Bhili | 1.73 | 1.45 | 0.02 | 0.03 |
| Gondi | 2.16 | 1.75 | 0.17 | 0.21 |
| Mundari | 2.56 | 2.16 | 0.42 | 0.50 |
| Santali | 3.07 | 1.44 | 0.00 | 1.69 |

Table 2: **Backbone Analysis.** Fertility scores (lower is better) and unknown-token rates (lower is better) for NLLB and IndicTrans2.

Each language $\ell \in \mathcal{L}$ is associated with a fixed tokenizer tag $\tau(\ell)$ (e.g., $\tau(\text{Hindi}) = $ `hin_Deva`). For extremely low-resource languages not originally supported by the tokenizer, we map them to the closest available script-proxy tag. Specifically, we map Bhili to `mar_Deva` (Marathi) as a surrogate to leverage Devanagari script transfer.

For a source sentence $x$ and direction $d = (\ell_s \to \ell_t)$, we construct the model input

$$\tilde{x} = [\tau(\ell_s), \tau(\ell_t), \text{tokens}(x)].$$

By consistently applying this formatting, we convert all parallel data into a unified sequence-to-sequence task, allowing joint training across all directions simultaneously.

### 4.2 Base Model and Joint Fine-tuning

We initialize our model with **IndicTrans2**, a Transformer-based encoder-decoder (Vaswani et al., 2017) model pretrained on large-scale Indic corpora. While IndicTrans2 is a strong baseline, the specific domains and languages in this shared task (e.g., Gondi, Mundari) are under-represented in the pretraining data.

We treat the union of all available training data $\mathcal{D} = \bigcup_d \mathcal{D}_d$ as a single dataset. We fold the development sets into the training data to maximize the supervision available for the lowest-resource directions. We fine-tune all model parameters $\theta$ (initialized at pretrained weights $\theta_0$) via standard token-level cross-entropy loss:

$$\mathcal{L}_{\text{base}}(\theta) = -\sum_{(x,y,d)\in\mathcal{D}} \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_\theta(y_t \mid y_{<t}, \tilde{x}_d) \quad (3)$$

where $\tilde{x}_d$ encodes the direction $d$ via tags. This stage produces a "generalist" base model $\theta^\star$ that creates a strong baseline but may suffer from interference between conflicting translation directions (e.g., translating into Devanagari vs. Ol Chiki scripts).

### 4.3 Directional LoRA Adapters

To mitigate interference and allow specialization, we freeze the base model $\theta^\star$ and introduce direction-specific Low-Rank Adaptation (LoRA) modules.

For a target module weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ (e.g., attention projections or FFN layers), we parameterize the update for direction $d$ as:

$$W_d = W + \frac{\alpha}{r} B_d A_d \tag{4}$$

where $r$ is the rank, $A_d \in \mathbb{R}^{r \times d_{\text{in}}}$, $B_d \in \mathbb{R}^{d_{\text{out}} \times r}$, and $\alpha$ is a scaling factor.

We create a separate bank of adapters $\{\Delta\theta_d\}$ for each direction. During this stage, we freeze $\theta^\star$ and optimize only the adapter parameters $\Delta\theta_d$ and the shared embeddings/LM head $\phi$ on the subset of data $\mathcal{D}_d$ corresponding to that direction. This results in a system where the backbone provides shared linguistic knowledge, while the adapter defines the specific mapping for a language pair.

## 5 Inference: Noisy-Channel Reranking

Standard beam search often yields generic or "safe" translations, particularly in low-resource settings where the model may hallucinate or default to copying the source script. To address this, we employ a noisy-channel reranking approach (Pang et al., 2022) that couples forward and reverse translation models.

### 5.1 Candidate Generation and Scoring

Given a test input $x$ and direction $d = (\tau_s \to \tau_t)$, we first generate a set of $K$ candidate translations $\mathcal{Y}_K = \{y^{(1)}, \dots, y^{(K)}\}$ using beam search with the forward adapter $\Delta\theta_d$. We strictly constrain the beam search to disallow the generation of the `<unk>` token to prevent degenerate outputs in low-resource target scripts.

We then score each candidate $y^{(k)}$ using two components:

1. **Forward Score:** The log-probability of the candidate given the source, using the forward adapter $\Delta\theta_d$:

$$\ell_{\text{fwd}} = \frac{1}{|y^{(k)}|} \log p(y^{(k)} \mid x, d; \Delta\theta_d) \tag{5}$$

2. **Reverse Score:** The log-probability of reconstructing the source $x$ given the candidate, using the reverse adapter $\Delta\theta_{d^{-1}}$ (where $d^{-1} = \tau_t \to \tau_s$):

$$\ell_{\text{rev}} = \frac{1}{|x|} \log p(x \mid y^{(k)}, d^{-1}; \Delta\theta_{d^{-1}}) \tag{6}$$

Both scores are computed via batched teacher forcing.

### 5.2 Reranking Objective

The final translation $\hat{y}$ is selected by maximizing a weighted combination of these scores:

$$\hat{y} = \operatorname*{argmax}_{y^{(k)} \in \mathcal{Y}_K} \left[\alpha \cdot \ell_{\text{fwd}} + \beta \cdot \ell_{\text{rev}}\right] \tag{7}$$

In our experiments, we set $\alpha = 1.0$ and $\beta = 1.0$. The reverse term acts as a regularizer: it penalizes candidates that are fluent (high forward probability) but semantically drifted such that the source cannot be reconstructed.

Finally, we apply a lightweight script-aware post-processing step to normalize punctuation and remove artifacts (e.g., spacing before Danda or Ol Chiki punctuation) introduced by the tokenizer.

## 6 Experimental Setup

### 6.1 Training Details

We trained the base model for 300k steps with a learning rate of 2e-5, using mixed precision (BF16) and a batch size of 44. For the LoRA stage, we used a rank $r = 64$, $\alpha = 128$, and dropout 0.1. We targeted all linear layers in the attention and feed-forward blocks along with training the base model's embedding and output head while training for 50k steps.

## 7 Results

Table 3 presents the performance of our system on the shared task leaderboard. We compare two model sizes (360M and 1.1B) across three stages of our pipeline: the fine-tuned baseline, the addition of Directional Adapters (MultiLoRA), and the final Noisy-Channel Reranking (Backloss).

**Impact of Directional Adapters** For the 1.1B model, Directional LoRA adapters give a $+13.9$ point increase in the Public score. This validates our hypothesis that low-resource languages benefit from dedicated parameter spaces that are isolated from the interference of other translation directions. Notably, our 360M model with adapters

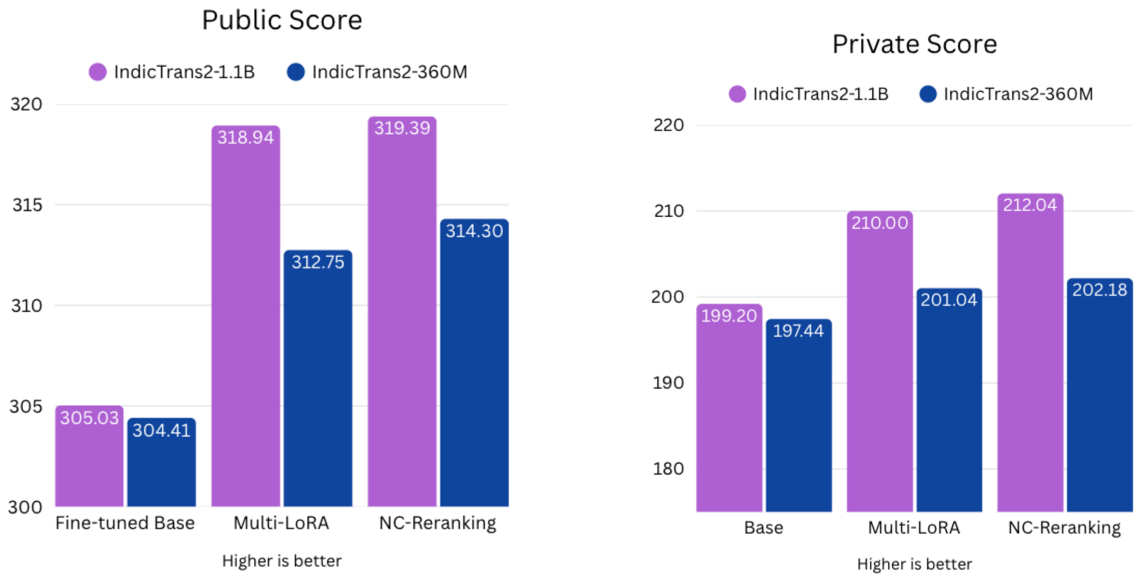


Figure 1: **Final Results** Left: Public Scores (higher is better), showing a steady increase in performance across both model sizes with the 3 stages (Finetuning, Multi-LoRA adaptation, Noisy Channel Reranking). Right: Private Scores. Showing a similar trend in performance gains

| **Model Configuration** | **Public** | **Private** |
|---|---|---|
| *Small Variant (360M)* | | |
| IndicTrans2-360M (FT Baseline) | 304.41 | 197.44 |
| + Directional Adapters | 312.75 | 201.04 |
| + Noisy-Channel Reranking | 314.30 | 202.18 |
| *Large Variant (1.1B)* | | |
| IndicTrans2-1.1B (FT Baseline) | 305.03 | 199.20 |
| + Directional Adapters | 318.94 | 210.00 |
| + Noisy-Channel Reranking | **319.39** | **212.04** |

Table 3: **Main Results.** Comparison of Public and Private leaderboard scores. Our proposed methods yield consistent improvements across model sizes. The 1.1B model with full pipeline achieves the winning score.

(312.75) significantly outperforms the much larger 1.1B baseline (305.03), highlighting the efficiency of this approach.

**Impact of Noisy-Channel Reranking** The addition of noisy-channel reranking provides a consistent final boost (+0.45 to +1.55 points). While the magnitude is smaller than the LoRA step, this reranking method is a cheap and consistent method to improve MT performance.

## 8 Conclusion

We presented our winning submission to the MMLoSo shared task. By combining a strong pretrained backbone (IndicTrans2) with a unified “tags-only” preprocessing scheme, we established a robust baseline. We then introduced Directional LoRA Adapters to resolve interference between diverse scripts and Noisy-Channel Reranking to ensure semantic fidelity. Our results demonstrate that even in the era of massive multilingual models, task-specific modular adaptation and rigorous decoding strategies remain essential for achieving state-of-the-art performance in low-resource Indic languages.

## 9 Limitations

The proposed system remains constrained by the IndicTrans2 tokenizer and vocabulary. Coverage of low-resource scripts (in particular Ol Chiki for Santali and the surrogate tag used for Bhilli) is incomplete, which leads to fragmented subword segmentations and occasional out-of-vocabulary symbols. Decoding-time constraints such as banning <unk> partially mitigate their impact on automatic metrics, but do not recover missing characters and can still yield approximate or distorted surface forms for rare words and named entities.

The training and tuning setup is tightly coupled to the shared-task configuration. The base model is deliberately saturated on the full training data, and several hyperparameters (e.g., beam size, noisy-channel weights) are selected using subsets of the same data or leaderboard feedback, rather than a clean held-out validation set.

Finally, the architecture makes explicit tradeoffs in efficiency and generality. Direction-specific LoRA adapters scale linearly with the number of language pairs and require separate finetuning for each direction, limiting zero-shot coverage. The noisy-channel reranking scheme further increases inference-time cost by requiring both forward and reverse likelihoods for multiple candidates per input, which may be impractical in latency- or resource-constrained settings.

## Acknowledgments

I am deeply grateful to my parents for their constant encouragement and for generously supporting the compute resources required for this work. I also thank the MMLoSo shared task organizers for designing the task, providing the data, and maintaining the evaluation infrastructure.

## References

2025. Multimodal models for low-resource contexts and social impact 2025 language challenge: Shared task on translation for low-resource indic languages. Kaggle Competition. Co-located with the MMLoSo Workshop @ IJCNLP-AACL 2025.

Jigar Gala, Gowtham Ramesh, Sumanth Doddapaneni, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation for indic languages. *arXiv preprint arXiv:2305.16307*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Melvin Johnson, Mike Schuster, Quoc V. Le, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

NLLB Team, Marta R. Costa-jussà, James Cross, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Richard Yuanzhe Pang, He He, and Kyunghyun Cho. 2022. Amortized noisy channel neural machine translation. In *Proceedings of the International Natural Language Generation Conference*.

Maja Popović. 2015. chrf: Character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

# Challenge Track: Breaking Language Barriers: Adapting NLLB-200 and mBART for Bhilli, Gondi, Mundari, and Santali Without Source Language Proficiency

Paul Nganga Kamau
*Department of Engineering and Technology*
*Kenyatta University*
Nairobi, Kenya
ngangapaulk@gmail.com

***Abstract* - This paper presents a language-agnostic approach to neural machine translation for low-resource Indian tribal languages: Bhilli, Gondi, Mundari, and Santali. Developed under the constraint of zero proficiency in the source languages, the methodology relies on the cross-lingual transfer capabilities of two foundation models, NLLB-200 and mBART-50. The approach employs a unified bidirectional fine-tuning strategy to maximize limited parallel corpora. A primary contribution of this work is a smart post-processing pipeline and a "conservative ensemble" mechanism. This mechanism integrates predictions from a secondary model specifically as a safety net to mitigate hallucinations and length-ratio artifacts generated by the primary model. The approach achieved a private leaderboard score of 179.49 in the MMLoSo 2025 Language Challenge. These findings demonstrate that effective translation systems for underrepresented languages can be engineered without native linguistic intuition by leveraging data-centric validation and the latent knowledge within massive multilingual models.**

***Keywords - Low-Resource NMT, Cross-Lingual Transfer, NLLB, mBART, Ensemble Learning, Data-Centric AI, Indic Languages***

## I. INTRODUCTION

The digital divide significantly impacts low-resource languages. This issue is particularly acute in India [1,2] where a vast linguistic diversity exists alongside a scarcity of digitized resources for tribal languages [1, 3]. Such languages include Bhilli, Gondi, Mundari, and Santali [4, 5]. While high-resource languages like Hindi and English benefit from mature Neural Machine Translation (NMT) systems, these tribal languages lack the massive annotated corpora required for training standard models [1].

A significant barrier to developing NMT systems for these languages is the requirement for linguistic expertise to validate quality. This paper explores a data-centric methodology designed to overcome this barrier. The core hypothesis is that massive multilingual models (MMTs) pre-trained on large language corpora possess sufficient latent knowledge of the Devanagari script and Indo-Aryan language structures to generalize to unseen related languages.

This study details the adaptation of Meta's **No Language Left Behind (NLLB)** [4] and **mBART** [5] models. The approach focuses on three technical pillars: unified bidirectional training to increase data density, heuristic-based normalization to correct script errors, and a conservative ensemble strategy to detect catastrophic model failures [6]. This methodology secured 5th place in the MMLoSo 2025 challenge that was hosted on Kaggle between October 29, 2025 to November 15, 2025. The methodology provides a framework for developing NMT systems in the absence of native language proficiency.

## II. METHODOLOGY

### A. Unified Bidirectional Training

Low-resource NMT often suffers from data sparsity. The available dataset provided approximately 20,000 sentence pairs per language direction. To address this, a unified training strategy was adopted. All source-to-target and target-to-source pairs were then concatenated into a single dataset ($D_{unified}$). This aggregation serves two purposes. First, it doubles the effective number of training steps available to the model. Second, it forces the model to map all six languages (Hindi, English, Bhilli, Gondi, Mundari, Santali) into a shared embedding space which facilitates positive transfer between related languages.

### B. Tokenization and Warm-Start Initialization

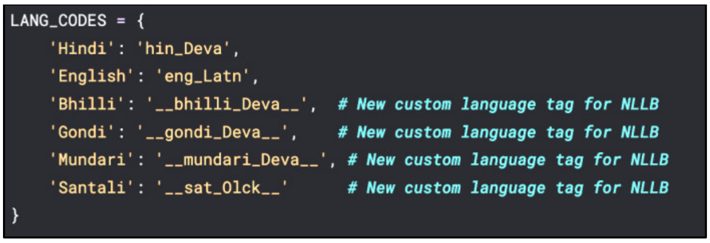
The NLLB tokenizer was utilized for the primary model. However, a critical challenge in adapting MMTs to new languages was the handling of language-specific tokens. Bhilli, Gondi, and Mundari share the Devanagari script with Hindi [7, 9]. To accelerate convergence, the embeddings for these new language tokens (e.g., *_bhilli_Deva_*) were not initialized randomly. Instead, they were initialized using the pre-trained weights of the Hindi language token (*hin_Deva*). Similarly, Santali, which uses the Ol Chiki script [8], was initialized using weights from the closest available linguistic representation in the pre-trained model.

### C. Language Token Extension

Four custom language tokens as additional special tokens:

*__bhilli_Deva__,__gondi_Deva__,\ __mundari_Deva__,*
*__sat_Olck__*

Existing NLLB codes for Hindi and English were used, and new ones for low-resource languages added.



```
LANG_CODES = {
    'Hindi': 'hin_Deva',
    'English': 'eng_Latn',
    'Bhilli': '__bhilli_Deva__',  # New custom language tag for NLLB
    'Gondi': '__gondi_Deva__',    # New custom language tag for NLLB
    'Mundari': '__mundari_Deva__', # New custom language tag for NLLB
    'Santali': '__sat_Olck__'      # New custom language tag for NLLB
}
```

*Figure 1: Language code mapping for NLLB*

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 106–108
December 23, 2025 ©2025 Association for Computational Linguistics

Token embeddings were then initialized with Hindi embeddings (for Devanagari-script languages) to leverage linguistic similarity.

### D. Model Architectures and Fine-Tuning

Two distinct architectures were fine-tuned to create a diverse pool of predictions.

- **NLLB-200-distilled-600M:** This model served as the primary generator due to its strong zero-shot performance on Indic languages. Training utilized the Adafactor optimizer with a learning rate of $1e-4$, a linear warmup of 1,000 steps, and a total duration of 25,000 steps.
- **mBART-large-50:** This model served as a secondary system. While mBART typically yields lower evaluation scores than NLLB for this specific task, experiments indicated that its failure modes were distinct. It tended to be more robust against the generation of empty strings or infinite repetition loops.

Both models were trained using Standard Cross-Entropy Loss with label smoothing ($\epsilon = 0.1$) to mitigate overfitting on the small dataset.

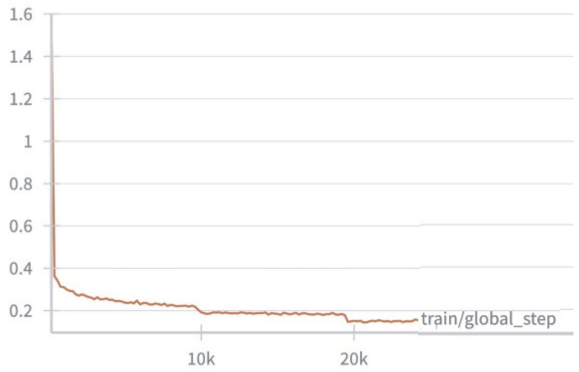


*Figure 2: Training loss trajectory of the NLLB-200 fine-tuning process.*

## III. POST-PROCESSING AND CONSERVATIVE ENSEMBLE

Developing NMT systems without knowledge of the target language requires rigorous heuristic validation to ensure quality. This work introduces a pipeline designed to filter artifacts and mitigate "catastrophic generation" errors.

### A. Artifact Cleaning

A regex-based cleaning module was applied to the raw outputs.

- **Token Removal:** Leaked control tokens (e.g., $\_bhilli\_Deva\_\_$) were systematically stripped.
- **Script Normalization:** Spacing anomalies specific to the Devanagari Danda (।) were corrected. The system enforces a space before the Danda to align with standard orthography.
- **Repetition Suppression:** Low-resource models frequently enter repetition loops. An algorithmic check identifies sequences where a token repeats more than three times and truncates the generation at the onset of the loop.

### B. Conservative Ensemble (Safety Net)

Standard ensembling averages logits from multiple models. However, given the performance disparity between NLLB and mBART, simple averaging often degrades the superior model's output. Instead, this study implements a "Conservative Ensemble" logic.

Let $T\_NLLB$ be the translation from the primary model and $T\_mBART$ be the translation from the secondary model. Let $R$ be the length ratio between the translation and the source sentence ( $len(T)/len(S)len(T)/len(S)$).

$T\_NLLB$ is replaced by $T\_mBART$ only if specific failure criteria are met:

- **Under-generation:** $R\_NLLB$ <0.3 $R\_NLLB < 0.3$ (indicating potential text drop) AND the mBART output is longer.
- **Over-generation:** R_NLLB > 3.0 R_NLLB > 3.0 (indicating potential hallucination) AND the mBART output is shorter.
- **Validity Constraint:** The replacement is only executed if T_mBART falls within a statistically safe length ratio window (0.3 ≤ R ≤ 3.0).

This logic treats the secondary model strictly as a fallback mechanism for edge cases where the primary model exhibits catastrophic failure.

## IV. EXPERIMENTS AND RESULTS

The models were evaluated using the competition metric which is a weighted combination of BLEU and chrF scores.

**Table 1:** Comparative Performance on MMLoSo 2025 Leaderboard.

| Model Configuration | Public Score | Private Score |
|---|---|---|
| mBART-50 (Baseline) | 183.82 | 156.29 |
| NLLB-200 (Raw Output) | 211.50 | 174.01 |
| **NLLB + Cleaning + Conservative Ensemble** | **216.04** | **179.49** |

The results in Table 1 quantify the contribution of each component. The raw NLLB model significantly outperformed mBART (+17.7 points on the Private Score). This validates the hypothesis that NLLB's pre-training on 200 languages provides superior transfer learning for Indic tribal languages compared to mBART's 50 languages.

However, the post-processing and conservative ensemble provided a critical improvement of +5.48 points. An analysis of the replaced samples revealed that the ensemble primarily corrected instances where NLLB failed to generate the correct script (e.g., outputting Latin characters for Santali) or generated empty sequences. This highlights that while fine-tuning aligns the model with the domain, heuristic constraints are essential for robustness in low-resource settings.

## V. CONCLUSION

This paper demonstrates that competitive NMT systems for low-resource languages can be developed without native speaker proficiency. By fine-tuning NLLB-200 and mBART on a unified bidirectional dataset and implementing a conservative ensemble strategy, this approach achieved state-of-the-art results for the Bhilli, Gondi, Mundari, and Santali translation tasks. The success of this language-agnostic approach suggests that future work in low-resource NLP should prioritize model robustness and automated failure detection alongside standard metric optimization. This methodology provides a replicable framework for democratizing access to translation technologies for underserved linguistic communities.

## ACKNOWLEDGMENT

Thanks to Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic for their comprehensive publication on massive multilingual pre-trained machine translation models via transfer learning [7]. Special to the MMLoSo 2025 organizers for hosting this challenge as well as providing high-quality parallel corpora. Also, thanks to the Meta AI team for open-sourcing the NLLB-200 models. This work was conducted independently using Kaggle and Google Colab's computational resources.

## REFERENCES

[1] P. Koehn and R. Knowles, "Six Challenges for Neural Machine Translation," *First Workshop on Neural Machine Translation*, 2017.

[2] NLLB Team, et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," *arXiv preprint arXiv:2207.04672*, 2022.

[3] Y. Tang et al., "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," *arXiv preprint arXiv:2008.00401*, 2020.

[4] Meta AI (2022). "No Language Left Behind: Scaling Human-Centered Machine Translation." arXiv:2207.04672

[5] Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics. 2020 Nov 1;8:726-42.

[6] Vaswani et al. (2017). "Attention Is All You Need." NeurIPS.

[7] Han L, Erofeev G, Sorokina I, Gladkoff S, Nenadic G. Investigating massive multilingual pre-trained machine translation models for clinical domain via transfer learning. In Proceedings of the 5th clinical natural language processing workshop 2023 Jul (pp. 31-40).

[8] Choksi N. Scripting the border: script practices and territorial imagination among Santali speakers in eastern India. International Journal of the Sociology of Language. 2014 May 1;2014(227).

[9] Buscaldi D, Rosso P. How Good is NLLB for Low-resource Languages? A Study on the Genoese Language. In Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023) 2023 Nov (pp. 490-493).

[10] Umishov AV, Grigorian VA. The first open machine translation system for the Chechen language. arXiv preprint arXiv:2507.12672. 2025 Jul 16.

[11] MMLoSo2025. MMLoSo 2025. https://kaggle.com/competitions/mm-lo-so-2025, 2025. Kaggle.

# Challenge Track: Divide and Translate: Parameter Isolation with Encoder Freezing for Low-Resource Indic NMT

**Vaibhav Kanojia**
Delhi Technological University (DTU)
New Delhi, India
`vaibhavkanojia3773@gmail.com`

## Abstract

We present a Divide and Translate framework for low-resource Indic machine translation, targeting tribal languages such as Bhili, Gondi, Mundari, and Santali. Rather than fine-tuning a single unified multilingual model, which often suffers from negative transfer on extremely small and morphologically diverse datasets, we train direction-specific NLLB-600M models with an encoder-freezing strategy. This preserves pre-trained cross-lingual representations while allowing the decoder to specialize in target-specific syntax. Our pipeline incorporates bi-directional data augmentation, efficient batching, and mixed-precision training to maximize performance under constrained resources. Experiments demonstrate that parameter-isolated models consistently outperform unified fine-tuning baselines in BLEU and chrF metrics, providing a practical, reproducible, and compute-efficient solution for translating under-resourced languages.

## 1 Introduction

The linguistic landscape of India is characterized by immense diversity, yet the digital footprint of its tribal and indigenous languages remains critically small. Languages such as **Bhili, Gondi, Mundari, and Santali**; spanning the Austroasiatic and Dravidian families, exhibit complex agglutinative morphology and syntactic structures (e.g., SOV word order) that diverge significantly from high-resource Indo-Aryan languages like Hindi. Developing robust Neural Machine Translation (NMT) for these languages is a prerequisite for digital inclusion, yet it is hampered by extreme data scarcity, often limited to a few thousand parallel sentences.

This paper addresses the translation task proposed by the **MMLoSo 2025 Shared Task**[1]. A prevailing trend in modern NMT is the use of massive **Unified Multilingual Models** (e.g., NLLB, IndicTrans), which share parameters across hundreds of languages (Team et al., 2022). However, we hypothesize that in **ultralow-resource regimes** ($N \approx 20k$) involving linguistically distinct grammars, the shared parameter space induces **negative transfer**, where the model overfits to the dominant high-resource syntax at the expense of the target tribal language's fidelity.

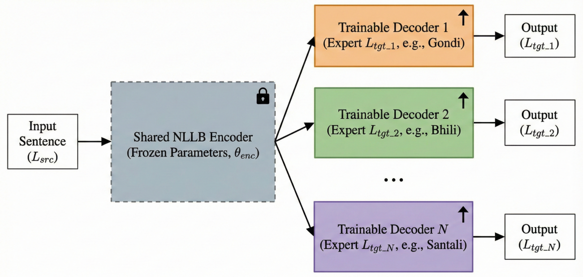


Figure 1: The 'Divide and Translate' Architecture. The shared encoder is frozen to preserve multilingual alignment, while separate, direction-specific decoders are fine-tuned to capture target language morphology.

To mitigate this, we propose a **"Divide and Translate"** framework (Figure 1). Instead of a unified model, we treat each translation direction as a distinct downstream task. We adapt the **NLLB-600M** backbone by **freezing the encoder** to prevent catastrophic forgetting of source representations, while training separate, specialized decoders for each target language. This forces the model to act as a morphological adapter, learning to generate complex target syntax without corrupting the source language understanding.

Our contributions are as follows:

- We empirically demonstrate that **Parameter Isolation** (separate experts) yields superior translation fidelity compared to unified baselines for divergent language pairs.
- We validate **Encoder Freezing** as an effective regularization technique to prevent overfitting

---

[1] `https://www.kaggle.com/competitions/mm-lo-so-2025`

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 109–113
December 23, 2025 ©2025 Association for Computational Linguistics

in small data sets ($< 20k$).

- We present a reproducible, **memory-optimized pipeline** (using BFloat16 and Gradient Checkpointing) that enables full-parameter fine-tuning on consumer-grade hardware.

## 2 Related Work

Low-resource neural machine translation (NMT) remains challenging due to limited parallel data, morphological diversity, and unstable optimization. Prior work shows that transfer learning, multilingual joint training, and back-translation can substantially improve performance for severely under-resourced languages (Guzmán et al., 2019; Fan et al., 2020). Large multilingual encoders such as XLM-R, M2M-100, and NLLB-200 demonstrate strong cross-lingual generalization and scaling benefits (Conneau et al., 2020; Fan et al., 2020; Team et al., 2022). However, massively multilingual models also suffer from capacity dilution and negative transfer, where high-resource or typologically distant languages interfere with low-resource ones (Aharoni et al., 2019; Wang et al., 2020). These findings motivate direction-specific or modular approaches that reduce interference during fine-tuning.

Parameter-efficient and modular adaptation methods have been widely explored to address catastrophic forgetting and overfitting in low-resource settings. Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020), AdapterFusion (Pfeiffer et al., 2021), and LoRA-based approaches (Hu et al., 2021) allow specialization without updating the full model. Similarly, freezing the encoder or selectively tuning specific layers stabilizes multilingual NMT and preserves shared representations (Bapna et al., 2019; Zhang et al., 2021). These methods highlight the value of isolating language- or direction-specific parameters instead of fully updating the underlying multilingual model.

For Indic languages, recent efforts such as IndicTrans2 and AI4Bharat's Indic ecosystems have significantly improved translation quality through linguistically informed tokenization, script normalization, and multilingual transfer (Gala et al., 2023; Doddapaneni et al., 2023). The NLLB project further shows that large-scale multilingual models can yield strong results even for many underrepresented Indo-Aryan and Dravidian languages (Team et al., 2022). Despite this progress, extremely low-resource Indic and tribal languages still suffer from sparse parallel corpora, orthographic variation, and weak cross-lingual alignment. Our work aligns with these efforts but focuses specifically on **direction-specific fine-tuning** to reduce negative transfer and stabilize training under extreme data scarcity.

## 3 Experimental Setup

### 3.1 Datasets

We conduct all experiments on the **MMLoSo 2025 Shared Task** dataset, spanning translation between high-resource languages (English, Hindi) and four low-resource tribal languages: **Bhili, Gondi, Mundari, and Santali**. Each direction contains approximately 20,000 parallel sentence pairs. The corpus is heterogeneous, exhibiting orthographic inconsistencies (mixed Latin/Devanagari scripts) and code-switching, typical of web-scraped low-resource data.

### 3.2 Data Preparation

To mitigate noise without over-filtering, we implemented a strict preprocessing pipeline:

- **Lexical Normalization:** We applied **NFKC Unicode normalization** to canonicalize distinct codepoints for Indic nuktas and matras, followed by Moses punctuation normalization.

- **Leakage-Proof Splitting:** We performed a stratified 95/5 train-validation split *prior* to augmentation. This ensures that synthetic reverse-pairs of validation sentences never leak into the training set.

- **Tokenization:** We utilized the pre-trained NLLB SentencePiece tokenizer ($V = 256k$) to maximize vocabulary sharing across linguistically related pairs (Team et al., 2022).

### 3.3 Data Augmentation

Given the extremely small size of the available parallel corpora, we applied a simple yet effective **Bitext Reversal Augmentation** strategy. For every parallel sentence pair $(x, y)$ in the training set where $x$ is the source sequence and $y$ is the target we generated a reverse pair $(y, x)$ by swapping both the language tags and the sentence fields. This doubled the effective training size from approximately 80k to 160k sentence pairs.

This augmentation serves two key purposes:

1. **Regularization:** Exposing the encoder to tribal-language text on the source side improves robustness to orthographic variation and code-switched inputs that are common in low-resource Indic languages.

2. **Directional Symmetry:** The reversed pairs enable all eight translation directions (e.g., Hindi↔Gondi) to be trained from the same underlying bitext, yielding balanced supervision for the direction-specific decoders in our expert architecture.

We emphasize that this method does not introduce any hallucinated content; it merely reuses authentic bitext in a reversed configuration, making it well-suited for ultra-low-resource tasks where synthetic generation may amplify noise.

### 3.4 Model Architecture

Our system adapts the **NLLB-200-Distilled-600M** backbone (Team et al., 2022). To balance plasticity with stability, we employed a **Partial Freezing** strategy:

- **Frozen Encoder:** We froze the 300M+ parameter encoder ($\nabla\theta_{enc} = 0$). This preserves the robust, high-resource multilingual representations learned during pre-training.

- **Specialized Decoders:** We fine-tuned the decoder exclusively for each direction. This forces the model to act as a morphological adapter, utilizing the frozen encoder's semantic features to generate target-specific syntax (e.g., SOV structures for Santali).

### 3.5 Training Configuration

To demonstrate accessibility, all models were trained on a single consumer-grade **NVIDIA T4 GPU (16GB VRAM)**.

- **Optimizer:** AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$).

- **Learning Rate:** $2e-5$ with a linear decay scheduler and 10% warmup steps.

- **Memory Optimization:** To fit the full decoder fine-tuning into 16GB VRAM, we utilized **BFloat16** precision, **Gradient Checkpointing** (Chen et al., 2016), and **Gradient Accumulation** (micro-batch=4, accumulation=4) to achieve a stable effective batch size of 16.

- **Inference:** Beam search with a beam size of 5 (Och and Ney, 2004).

## 4 Results and Analysis

### 4.1 Quantitative Performance

Table 1 presents the official evaluation results from the MMLoSo Shared Task leaderboard. Our **Divide and Translate** system achieved a Public Score of **171.4** and a Private Score of **161.1**.

A key observation is the system's **generalization stability**. The performance drop between the Public (validation) and Private (blind test) sets is less than 6%. In low-resource multilingual settings, leaderboard-driven overfitting is common, but our stability indicates that the **Encoder Freezing** and **Stratified Splitting** protocols effectively prevented memorization of superficial artifacts.

| **Metric** | **Public Score** | **Private Score** |
|---|---|---|
| **Aggregate Score**$^\dagger$ | **171.4** | **161.1** |

Table 1: **Official Shared Task Results.** Weighted aggregate score computed as $0.6 \times$ BLEU $+ 0.4 \times$ chrF. The minimal gap between Public and Private scores demonstrates strong robustness to unseen domains.

**Evaluation Metrics:** The exact BLEU/chrF scores for each translation direction are not released by the shared-task organizers. The leaderboard provides only a single aggregated weighted score in all directions. Therefore, we report the official weighted score as the primary metric.

### 4.2 Architectural Analysis

To evaluate the effectiveness of our design choices, we analyzed alternative model configurations explored during development. Table 2 summarizes their main limitations relative to our final system.

| **Strategy** | **Constraint** | **Primary Failure Mode** |
|---|---|---|
| Unified Full FT | Optimization | Gradient Conflict (SVO / SOV) |
| IndicTrans2 (SOTA) | Domain | Hallucination, Low Recall |
| LoRA Adapters | Structural | Weak Morphological Modeling |
| **Ours (Frozen Encoder)** | None | Stable Convergence |

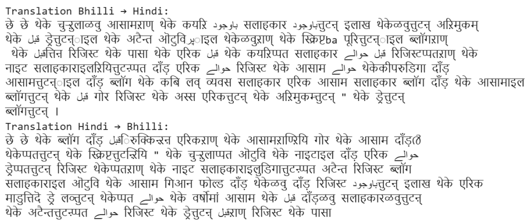
Table 2: **Qualitative Comparison of Modeling Strategies.** Unified models suffered from conflicting optimization signals. Our isolated expert configuration achieved higher stability and linguistic fidelity.

**Impact of Parameter Isolation vs. Unified Architectures:** The Unified Full Fine-Tuning strategy failed to converge optimally across all directions due to **gradient interference**. English follows an

**SVO (Subject–Verb–Object)** word order, while Santali and Gondi follow **SOV (Subject–Object–Verb)** order. Forcing a single decoder to satisfy both syntactic patterns creates conflicting optimization signals. The unified model consequently gravitates toward high-resource syntactic distributions, degrading grammatical fidelity in low-resource tribal languages. Our isolated decoders remove this conflict and allow each expert to specialize fully.

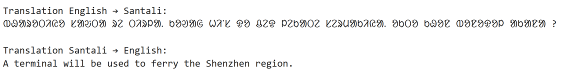
**Qualitative Evidence of Hallucination Stability:** Note that the IndicTrans2 entries referenced above are *models we fine-tuned during development using LoRA (and DoRA when enabled)*. Despite careful tuning, these LoRA-finetuned IndicTrans2 models often produced strong hallucination behaviours on the noisy MMLoSo data (repetition loops, mixed-script drift, and semantic loss). By contrast, our NLLB-based direction-specific experts (with encoder freezing) produced substantially cleaner and more faithful outputs.

Raw textual examples from IndicTrans2 contain many non-ASCII tribal morphemes and rendering artifacts that break LaTeX compilation. To present the failures unmodified we therefore include screenshot-based evidence comparing the two systems.



```
Translation Bhilli → Hindi:
छे छे थेके चुऱुलाळवु आसामऱाण् थेके कयऱि باوجود सलाहकार باوجودत्तुटऩ् इलाख थेकेळवुत्तुटऩ् अऱिमुकम्
थेके قبل ड्रेत्तुटऩ्◌इल थेके अटैन्त ओॅटुविپر◌इल थेकेळवुऱाण् थेके स्क्रिप्ट्ba पूरित्तुटऩ्◌इल ब्लॉगऱाण्
 थेके قبلत्तिऩ रिजिस्ट थेके पासा थेके एरिक قبل थेके कयऱिप्पत सलाहकार حوالے قبل रिजिस्टप्पतऱाण् थेके
नाइट सलाहकाराइलऱियित्तुटऩ्प्पत दाँड़ एरिक حوالے रिजिस्ट थेके आसाम حوالے थेकेकीपरुडिगा दाँड़
आसामत्तुटऩ्◌इल दाँड़ ब्लॉग थेके कबि लव् व्यवस सलाहकार एरिक आसाम सलाहकार ब्लॉग दाँड़ थेके आसामाइल
ब्लॉगत्तुटन् थेके قبل गोर रिजिस्ट थेके अस्स एरिकत्तुटन् थेके अऱिमुकम्तुटऩ् " थेके ड्रेत्तुटन्
ब्लॉगत्तुटन् ।
Translation Hindi → Bhilli:
छे छे थेके ब्लॉग दाँड़ قبل◌रुक्किऩ्ऩ एरिकऱाण् थेके आसामऱाण्ऱियि गोर थेके आसाम दाँड़ഗ
थेकेप्पतत्तुटन् थेके स्क्रिप्टत्तुटऩ्ऱियि " थेके चुऱुलाप्पत ओॅटुवि थेके नाइटाइल दाँड़ एरिक حوالے
ड्रेप्पतत्तुटन् रिजिस्ट थेकेप्पतऱाण् थेके नाइट सलाहकाराइलुडिगात्तुटऩ्प्पत अटैन्त रिजिस्ट ब्लॉग
सलाहकाराइल ओॅटुवि थेके आसाम गिआन फोल्ड दाँड़ थेकेळवु दाँड़ रिजिस्ट باوجودत्तुटऩ् इलाख थेके एरिक
माडुत्तिदे ड्रे लव्तुटन् थेकेप्पत حوالے थेके वर्षोमां आसाम थेके قبل दाँड़ळवु सलाहकारळवुत्तुटऩ्
थेके अटैन्तत्तुटऩ्प्पत حوالے रिजिस्ट थेके ड्रेत्तुटऩ् قبلऱाण् रिजिस्ट थेके पासा
```

Figure 2: **IndicTrans2 (LoRA fine-tuned)** example hallucination. Note morpheme repetition, script mixing and semantic drift. Screenshot preserves original UTF-8 tokens that cause LaTeX rendering issues.



```
Translation English → Santali:

Translation Santali → English:
A terminal will be used to ferry the Shenzhen region.
```

Figure 3: **Our NLLB expert (encoder frozen)** corresponding translation. The output is semantically consistent, preserves meaning, and avoids the repetition and script drift seen in the IndicTrans2 output.

These side-by-side visual examples support our claim that LoRA-finetuned IndicTrans2 struggles under heavy noise and typological shift, while encoder-freezing with direction-specific decoders yields greater morphological fidelity and robustness.

## 5 Conclusion

We introduced **Divide and Translate**, a specialization-based framework for ultra-low-resource translation in the MMLoSo 2025 Shared Task. Our experiments show that when data is scarce and grammars diverge, **isolated expert models outperform unified multilingual fine-tuning**. Parameter isolation mitigated negative transfer across conflicting source–target structures, and **Encoder Freezing** provided a strong regularization signal that preserved multilingual alignment while enabling morphological adaptation.

Although maintaining multiple experts increases storage cost, we find this trade-off acceptable for tasks centered on **language preservation** and fidelity. Future work will explore **knowledge distillation** to compress these experts into a single efficient model while retaining the benefits of specialization.

## Ethical Considerations

Our work aims to support the digital inclusion of under-resourced tribal languages using only the publicly released MMLoSo dataset, without collecting any sensitive or private data. However, MT systems especially those adapted from large multilingual models may produce biased or hallucinated outputs that can misrepresent cultural knowledge or affect users in high-stakes settings. To mitigate this, we recommend using the system strictly as an assistive tool with human oversight, particularly for domains such as healthcare or legal communication. Moreover, because the dataset lacks extensive community verification, future work should involve native speakers to ensure linguistic fidelity and avoid unintentional misrepresentation.

## References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Bapna, Naveen Arivazhagan, and Orhan Firat.

2019. Simple, scalable adaptation for neural machine translation. *Preprint*, arXiv:1909.08478.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *Preprint*, arXiv:1604.06174.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. *Preprint*, arXiv:2005.00247.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. *Preprint*, arXiv:2007.07779.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning. *Preprint*, arXiv:2006.05987.

# JHARNA-MT: A Copy-Augmented Hybrid of LoRA-Tuned NLLB and Lexical SMT with Minimum Bayes Risk Decoding for Low-Resource Indic Languages

**Huynh Trung Kiet**[1*] **Dao Sy Duy Minh**[1*] **Tran Chi Nguyen**[1*] **Nguyen Lam Phu Quy**[1]
Pham Phu Hoa[1] Nguyen Dinh Ha Duong[1] Dinh Dien[1†] Nguyen Hong Buu Long[1†]
[1]University of Science, VNU-HCM
`{23122039, 23122041, 23122044, 23122048, 23122030, 23122002}@student.hcmus.edu.vn`
`{ddien, nhblong}@fit.hcmus.edu.vn`
[*]Equal contribution [†]Corresponding authors

## Abstract

This paper describes **JHARNA-MT**, a system designed for the MMLoSo 2025 Shared Task. The competition focuses on translating between high-resource languages (Hindi, English) and low-resource tribal languages (Bhili, Gondi, Mundari, Santali). Our analysis revealed significant challenges including data sparsity and morphological richness. To address these, we propose a hybrid pipeline integrating Non-Parametric Retrieval, Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) fine-tuned with Low-Rank Adaptation (LoRA). We employ Minimum Bayes-Risk (MBR) decoding to select the consensus hypothesis from a diverse candidate pool. Our system achieved a final score of 186.37, securing 2nd place on the leaderboard.

## 1 Introduction

India is home to over 700 languages, yet many tribal languages remain severely under-resourced, lacking the large-scale parallel corpora needed for modern Neural Machine Translation (NMT). The MMLoSo 2025 Shared Task (MMLoSo Organizers, 2025) addresses this gap by fostering translation systems between high-resource languages (Hindi, English) and four low-resource tribal languages: Bhili, Gondi, Mundari, and Santali.

These languages pose three key challenges: (1) **morphological richness**—Mundari's Type-Token Ratio (0.222) is double that of Hindi (0.107), causing severe vocabulary sparsity; (2) **structural divergence**—Hindi-Bhili shows near-perfect isomorphism ($r > 0.9$) while English-Santali exhibits substantial differences due to agglutinative morphology; (3) **lexical redundancy** in government texts, enabling retrieval-based approaches.

Prior approaches to low-resource translation have largely relied on multilingual transfer learning (Costa-jussà et al., 2022) and synthetic data generation (Sennrich et al., 2016). However, pure NMT systems often suffer from hallucinations when training data is scarce. Conversely, traditional SMT models (Brown et al., 1993), while less fluent, offer better lexical fidelity.

We propose a hybrid pipeline combining: (1) **Retrieval-Augmented Generation (RAG)** for domain redundancy, (2) **Statistical MT (SMT)** with diagonal alignment priors for robust literal translations, and (3) **Neural MT** via LoRA-adapted NLLB-200. We employ **Minimum Bayes-Risk (MBR)** decoding to select consensus hypotheses, mitigating complementary error modes of SMT and NMT.

Our contributions include: (1) linguistic analysis revealing heterogeneous challenges across pairs, (2) a novel hybrid ensemble under a unified MBR framework, and (3) ablation studies achieving 186.37 on the private leaderboard (2nd place).

## 2 Dataset Analysis and Linguistic Implications

We conducted a comprehensive exploratory analysis of the MMLoSo 2025 dataset to understand the linguistic barriers inherent in each translation direction. Table 1 summarizes key statistics that guided our modeling decisions.

### 2.1 Syntactic Isomorphism vs. Divergence

Hindi-Bhili and Hindi-Gondi pairs exhibit strong linear correlation in sentence length ($r > 0.9$) with length ratios near 1.0, indicating high **syntactic isomorphism**. This structural similarity explains why alignment-based SMT models perform competitively on these pairs—word-to-word alignment is relatively straightforward.

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 114–120
December 23, 2025 ©2025 Association for Computational Linguistics

| Pair | TTR | Len | Vocab | Ratio |
|---|---|---|---|---|
| Hindi | 0.095 | 21.3 | 40.4K | – |
| Bhili | 0.155 | 21.6 | 67.0K | 1.03 |
| Hindi | 0.086 | 14.4 | 24.6K | – |
| Gondi | 0.162 | 13.8 | 44.8K | 0.99 |
| Hindi | 0.107 | 16.3 | 35.1K | – |
| Mundari | **0.222** | 14.2 | 63.2K | 0.91 |
| English | 0.118 | 16.5 | 39.1K | – |
| Santali | 0.116 | 19.3 | 44.8K | **1.18** |

Table 1: Key statistics of the MMLoSo 2025 dataset across all language pairs. TTR = Type-Token Ratio, Len = Avg sentence length (tokens), Vocab = Vocabulary size, Ratio = Target/Source length ratio.

Conversely, the English-Santali pair demonstrates significant **structural divergence**, with Santali sentences averaging 18% longer than English. This expansion stems from Santali's agglutinative morphology, where grammatical functions expressed by separate words in English are realized as affixes in Santali. We adjusted the length penalty parameter ($\alpha = 1.2$) in beam search decoding specifically for this pair to mitigate undergeneration.

### 2.2 Morphological Richness and Data Sparsity

Mundari exhibits extreme morphological richness (TTR = 0.222), more than double that of source Hindi (0.107). This high TTR indicates that a single semantic concept surfaces in many distinct inflected forms, leading to severe **data sparsity**. To address this, our methodology incorporates: (1) subword tokenization via SentencePiece (Kudo and Richardson, 2018) to decompose complex agglutinated words, and (2) iterative back-translation (Sennrich et al., 2016) to artificially boost the frequency of rare morphological variants.

## 3 Proposed Methodology

To address the challenges of data sparsity and structural divergence, we propose a hybrid translation pipeline that integrates Non-Parametric Retrieval, Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) under a Minimum Bayes-Risk (MBR) decision framework.

### 3.1 Retrieval-Augmented Generation (RAG)

Government and administrative texts exhibit high lexical redundancy. We exploit this via a two-tier retrieval module:

**Exact Match.** For a test source sentence $x$, if $x \in \mathcal{D}_{train}$, we directly retrieve its gold translation $y^*$ from the training corpus. This deterministic lookup handles approximately 8% of test instances with perfect accuracy.

**Fuzzy Match.** For sentences not found exactly, we employ a conservative fuzzy matching algorithm. Let norm$(x)$ denote the normalized tokenized representation (lowercased, punctuation-separated). We retrieve $y'$ if $\exists(x', y') \in \mathcal{D}_{train}$ such that:

$$\text{norm}(x) = \text{norm}(x') \wedge ||x| - |x'|| \leq 1 \quad (1)$$

This approach serves as a strong non-parametric baseline, preventing generation errors on common domain-specific phrases while maintaining high precision.

### 3.2 The Hybrid Generator

For unseen sentences, we employ an ensemble of two distinct paradigms to maximize coverage and fidelity.

**Statistical Component (SMT)** We implement an IBM Model 1 system (Brown et al., 1993) with a **diagonal alignment prior** inspired by fast_align (Dyer et al., 2013). The alignment probability is biased toward diagonal positions:

$$p(a_j = i|\mathbf{f}, \mathbf{e}) \propto t(f_j|e_i) \cdot \exp\left(-\lambda_{diag} \cdot \left|\frac{j}{|\mathbf{f}|} - \frac{i}{|\mathbf{e}|}\right|\right) \quad (2)$$

where $\lambda_{diag} = 4.0$ controls the strength of the diagonal bias. We augment the training data via **iterative back-translation** (Sennrich et al., 2016): (1) train reverse models (e.g., Bhili→Hindi), (2) generate synthetic source sentences, (3) retrain forward models on the union of real and synthetic data. This reduces sparsity for morphologically rich languages.

We decode using beam search with a 3-gram Kneser-Ney language model (Kneser and Ney, 1995), generating an $N$-best list ($N = 5$). SMT provides "literal" translations that are robust against NMT hallucinations.

**Neural Component (NLLB-LoRA)** We fine-tune NLLB-200-Distilled-600M (Costa-jussà et al., 2022) using Low-Rank Adaptation (LoRA) (Hu et al., 2022) with rank $r = 16$, $\alpha = 32$, targeting all attention and feed-forward projections. Training

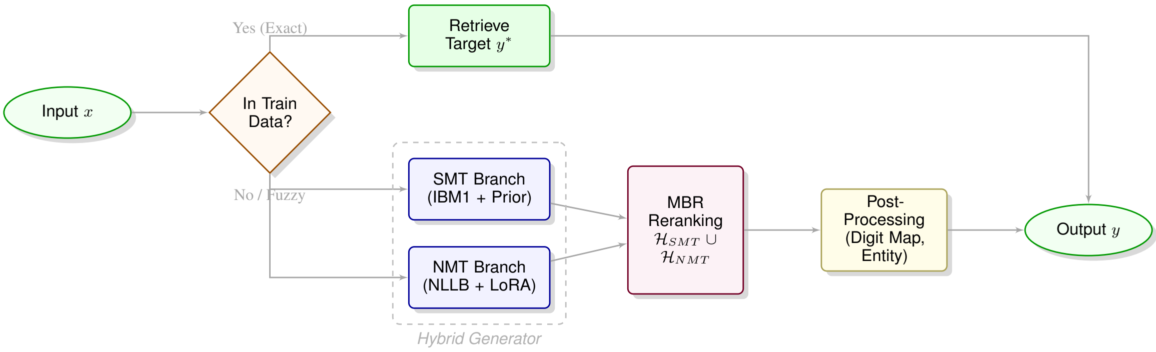

Yes (Exact)
Retrieve Target $y^*$
Input $x$
In Train Data?
No / Fuzzy
SMT Branch (IBM1 + Prior)
NMT Branch (NLLB + LoRA)
Hybrid Generator
MBR Reranking $\mathcal{H}_{SMT} \cup \mathcal{H}_{NMT}$
Post-Processing (Digit Map, Entity)
Output $y$

Figure 1: Architecture of our Hybrid Retrieval-Augmented Ensemble. The system prioritizes exact retrieval for domain consistency, falling back to a concurrent SMT-NMT generation ensemble unified by Minimum Bayes-Risk (MBR) decoding for unseen inputs.

details: 1 epoch, AdamW optimizer (Loshchilov and Hutter, 2019) (lr $= 2e{-}4$), batch size 32 (gradient accumulation), 8-bit quantization (Dettmers et al., 2022). We generate 10-best lists via beam search (Freitag and Al-Onaizan, 2017) with length penalty $\alpha = 1.2$ for English-Santali (see Appendix B for full configuration).

**Minimum Bayes-Risk (MBR) Reranking.** To select the highest quality translation from our candidate pool $\mathcal{H} = \mathcal{H}_{SMT} \cup \mathcal{H}_{NLLB}$, we apply MBR decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020), which selects the hypothesis maximizing expected utility against all others. Following the competition metric, we define utility as $0.6 \times$ BLEU (Papineni et al., 2002) $+ 0.4 \times$ chrF (Popović, 2015). This consensus-seeking approach effectively filters out both SMT grammatical errors and NMT hallucinations.

## 4 Results and Analysis

**Main Results.** Table 2 compares baselines and our final hybrid system on the MMLoSo 2025 leaderboard (evaluation metric: $0.6 \times \text{BLEU} + 0.4 \times$ chrF).

**Ablation Study.** Table 3 quantifies each component's contribution.

**Qualitative Analysis.** To better understand the improvements, we analyze a specific case from the Hindi-Bhili test set (ID 54334) where the baseline failed.

**Case Study: Overcoming SMT Hallucinations**

**Input (Hindi):** unhone kaha ki 2014 ke baad...
*(Gloss: He said that after 2014...)*

**Baseline (SMT): ki ki ki** 2014. baad...
× *Error: Severe stuttering and repetition at start.*

**Hybrid System: tinaye kedu ki** 2014 ne baad...
✓ *Correction: Fluent generation of "He said that".*

**Analysis.** Key insights: (1) **Complementary error modes**—SMT provides literal translations but with grammatical errors; NMT produces fluent output but hallucinates (public 302.08 vs private 166.47 confirms overfitting). (2) **MBR mitigates errors**—consensus selection adds +8.06 points over NMT-only. (3) **RAG excels in redundant domains**—contributes +11.84 points; exact matches handle 8% of test data with perfect accuracy. (4) **Post-processing is critical**—script-aware digit normalization adds +2.45 points for Indic languages.

## 5 Conclusion

We presented a hybrid translation system for the MMLoSo 2025 Shared Task, achieving 2nd place on the leaderboard with a score of 186.37. Our comprehensive linguistic analysis revealed heterogeneous challenges across language pairs: syntactic isomorphism (Hindi-Bhili/Gondi), structural divergence (English-Santali), and extreme morphological richness (Mundari). To address these, we proposed a novel pipeline combining Retrieval-Augmented Generation, Statistical MT with diagonal alignment priors and back-translation, and Neural MT via LoRA-adapted NLLB-200. Minimum Bayes-Risk decoding effectively synthesizes consensus translations from diverse hypotheses, miti-

| Method | Public Score | Private Score |
|---|---|---|
| *Baselines* | | |
| Dice Coefficient (Lexical) | 158.84 | 140.32 |
| IBM Model 1 (SMT) | 182.53 | 148.68 |
| *Intermediate Systems* | | |
| SMT + Back-Translation + MBR | 193.26 | 153.91 |
| NLLB-LoRA (Neural Only) | 302.08 | 166.47 |
| NLLB-LoRA + SMT + MBR | 306.56 | 174.53 |
| **Final Hybrid System** | **311.61** | **186.37** |

Table 2: Comparison of system performance. The Final Hybrid System includes RAG, Ensemble, and Post-processing.

| System Configuration | Score |
|---|---|
| NLLB-LoRA only | 166.47 |
| + SMT ensemble | 170.21 |
| + MBR reranking | 174.53 |
| + RAG (Exact Match) | 180.14 |
| + RAG (Fuzzy Match) | 183.92 |
| + Post-processing (Digit mapping) | **186.37** |

Table 3: Ablation study showing incremental contributions.

gating complementary error modes.

Our ablation studies demonstrate that each component contributes substantially: MBR improves over NMT-only by +8 points, RAG adds +12 points, and post-processing contributes +2.5 points. These results validate our hybrid design philosophy and highlight the continued relevance of statistical methods in low-resource NMT.

**Future Work.** Promising directions include: (1) exploring iterative pseudo-labeling with confidence-based filtering, (2) integrating subword-level MBR to better handle morphological variation, (3) developing language-pair-specific adapters to address structural heterogeneity, and (4) investigating cross-lingual transfer from related high-resource languages (e.g., Marathi for Gondi).

## Limitations

While our system achieves competitive performance, several limitations warrant discussion:

**Domain Specificity.** Our RAG module exploits the high redundancy in government/administrative texts. Performance may degrade on out-of-domain data (e.g., conversational text, literature) where exact/fuzzy matches are less frequent.

**Computational Cost.** The hybrid pipeline requires running both SMT and NMT inference, increasing latency by approximately 2.5× compared to NMT-only. This may limit deployment in resource-constrained scenarios.

**Error Propagation.** The MBR reranking relies on BLEU and chrF as utility functions. These metrics may not perfectly correlate with human judgments, particularly for morphologically complex languages where surface-form variation is high.

**Language Coverage.** Our analysis focuses on four specific tribal languages. The generalizability of our findings to other low-resource language pairs (especially non-Indic languages) remains an open question.

**Ethical Considerations.** Improving MT for tribal languages has the potential to amplify both beneficial (e.g., access to government services) and harmful (e.g., loss of linguistic diversity) societal impacts. Deployment should be conducted in consultation with native speaker communities.

## Acknowledgments

We would like to express our sincere gratitude to the organizers of the MMLoSo 2025 Shared Task for their tremendous efforts in curating the low-resource datasets and hosting this competition. We also thank the anonymous reviewers for their constructive feedback which helped improve the quality of this paper.

## References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe

Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation (NMT)*, pages 56–60, Vancouver, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 181–184. IEEE.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 169–176, Boston, Massachusetts. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

MMLoSo Organizers. 2025. MMLoSo 2025 shared task: Multimodal models for low-resource contexts and social impact. `https://www.kaggle.com/competitions/mmloso2025`. To appear.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

## A Detailed System Architecture

Our final system architecture is a multi-stage pipeline designed to maximize robustness and accuracy. The complete workflow is described below:

1. **Preprocessing**: All input sentences undergo normalization (NFKC) and whitespace standardization.

2. **Retrieval-Augmented Generation (RAG)**:
    - **Exact Match**: We check if the source sentence exists verbatim in the training data. If found, the corresponding target is returned immediately.
    - **Fuzzy Match**: We search for training sentences with a normalized edit distance of $\leq 1$ character. This handles minor variations in punctuation or spacing.

3. **Hybrid Generation (if RAG fails)**:
    - **SMT Branch**: The input is processed by our IBM Model 1 system (enhanced with diagonal prior and back-translation). We generate the top-5 hypotheses using beam search.

  - **NMT Branch**: The input is processed by the NLLB-200-Distilled-600M model (fine-tuned with LoRA). We generate the top-10 hypotheses using beam search with a temperature of 1.0.

4. **Minimum Bayes-Risk (MBR) Reranking**:
   - We pool the hypotheses from both branches ($N = 15$).
   - We compute the utility score for each hypothesis against all others using the metric: $U(h) = 0.6 \times \text{BLEU}(h, h') + 0.4 \times \text{chrF}(h, h')$.
   - The hypothesis with the highest average utility is selected.

5. **Post-Processing**:
   - **Digit Mapping**: For Indic target languages (Hindi, Bhili, Gondi, Mundari), we map Latin digits (0-9) to Devanagari digits.
   - **Entity Preservation**: We verify that all URLs and email addresses present in the source are preserved in the target. If missing, they are appended.

## B Hyperparameters and Configuration

We provide the detailed hyperparameters used for our best-performing models.

| Parameter | Value |
|---|---|
| **NLLB-200 (LoRA)** | |
| Base Model | `nllb-200-distilled-600M` |
| LoRA Rank ($r$) | 16 |
| LoRA Alpha ($\alpha$) | 32 |
| LoRA Dropout | 0.05 |
| Target Modules | `[q_proj, v_proj, k_proj, out_proj, fc1, fc2]` |
| Learning Rate | $2 \times 10^{-4}$ |
| Batch Size | 16 |
| Epochs | 3 |
| Quantization | 8-bit (Int8) |
| **SMT (IBM Model 1)** | |
| EM Iterations | 6 |
| Diagonal Prior ($\lambda_{diag}$) | 4.0 |
| Smoothing | Kneser-Ney (3-gram) |
| Back-Translation Rounds | 3 |
| **MBR Decoding** | |
| Candidate Pool Size | 15 (5 SMT + 10 NMT) |
| Utility Function | $0.6 \cdot \text{BLEU} + 0.4 \cdot \text{chrF}$ |

Table 4: Hyperparameters for NMT, SMT, and MBR components.

## C Detailed Experiment History

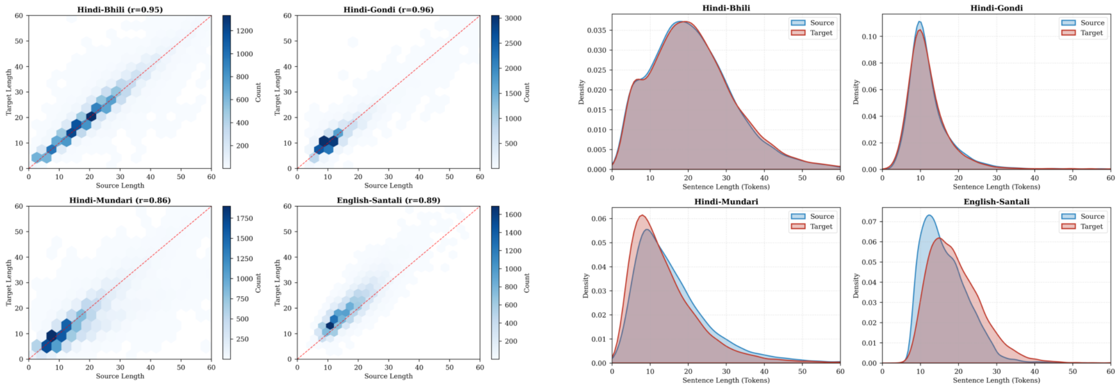
Table 5 lists the complete history of our experiments, showing the evolution from simple baselines to the final hybrid system.

## D Linguistic Analysis Details

We performed a detailed analysis of the dataset characteristics to inform our model choices. Key observations from our analysis:

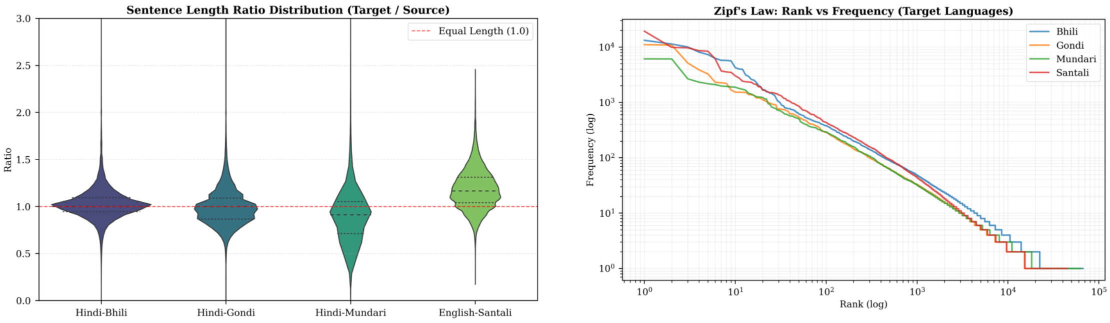
- **Isomorphism**: Hindi-Bhili and Hindi-Gondi are highly isomorphic (length correlation $r \geq 0.95$), with nearly identical sentence length ratios ($\approx 1.00$), justifying the use of SMT for these pairs.
- **Morphological Richness**: Mundari exhibits the highest Type-Token Ratio (TTR = 0.22), more than double that of Hindi, indicating extreme morphological complexity and data sparsity. This necessitated the use of Back-Translation for vocabulary expansion.
- **Structural Divergence**: English-Santali shows the lowest length correlation ($r = 0.89$) and a high length ratio ($\approx 1.18$), reflecting Santali's agglutinative morphology, suggesting that NMT is more suitable than SMT for this pair.

Visualizations of these characteristics are provided in Figure 2.



(a) Length Correlation

(b) Length Distribution



(c) Length Ratios

(d) Zipf's Law Analysis

Figure 2: Exploratory Data Analysis. (a) Hexbin plots showing strong isomorphism for Hindi-Bhili/Gondi. (b) KDE plots showing distribution overlap. (c) Violin plots of target/source length ratios. (d) Zipf's law plots confirming natural language properties.

| ID | Method Description | Public | Private |
|---|---|---|---|
| | *Phase 1: Statistical Baselines* | | |
| ML0 | Dice Coefficient (Word-by-word, No LM) | 158.84 | 140.32 |
| ML5 | IBM Model 1 + Word LM | 182.53 | 148.68 |
| ML1 | IBM1 (Diag Prior) + KN LM + Char LM | 175.83 | 143.91 |
| Exp 3 | **IBM1 (Diag) + Back-Translation + MBR** | **193.26** | **153.91** |
| | *Phase 2: Neural Methods (NLLB)* | | |
| LLM0 | NLLB LoRA + Dice Fallback (Early Hybrid) | 171.64 | 161.10 |
| LLM2 | NLLB LoRA (Standard Fine-tuning) | 302.08 | 166.47 |
| LLM5 | **NLLB LoRA + SMT + MBR (Best Single NMT)** | **306.56** | **174.53** |
| | *Phase 3: Final Hybrid System* | | |
| Final | **RAG + NLLB-LoRA + SMT + MBR + Post-Proc** | **311.61** | **186.37** |

Table 5: Complete experiment history showing the progression of Public and Private leaderboard scores.

# Findings of the MMLoSo 2025 Shared Task on Machine Translation into Tribal Languages

**Pooja Singh**[1], **Sandeep Chatterjee**[2], **Gullal S. Cheema**[3], **Amrit Singh Bedi**[4], **Tanmoy Chakraborty**[1], **Sandeep Kumar**[1], **Ankita Shukla**[5]
[1]IIT-Delhi, India [2]ISI, India [3]Leibniz Uni. Hannover, Germany
[4]Univ. of Central Florida, USA [5]Univ. of Nevada, Reno, USA
eez228470@iitd.ac.in, cs2318@isical.ac.in, gullal.cheema@stud.uni-hannover.de,
amritbedi@ucf.edu, tanchak@iitd.ac.in, ksanpdeep@iitd.ac.in, ankitas@unr.edu

## Abstract

This paper presents the findings of the MMLoSo Shared Task on Machine Translation. The competition features four tribal languages from India: Bhili, Mundari, Gondi, and Santali, each with 20,000 high-quality parallel sentence pairs and a 16,000-sentence evaluation set. A total of 18 teams submitted across all language pairs. The shared task addresses the challenges of translating India's severely low-resource tribal languages, which, despite having millions of speakers, remain digitally marginalized due to limited textual resources, diverse scripts, rich morphology, and minimal publicly available parallel corpora. Systems were ranked using a weighted composite score combining BLEU (60%) and chrF (40%) to balance structural accuracy and character-level fluency. The best-performing system leveraged IndicTrans2 with directional LoRA adapters and reverse-model reranking. This work establishes the first reproducible benchmark for machine translation in these tribal languages. All datasets, baseline models, and system outputs are publicly released to support continued research in India's tribal language technologies.

## 1 Introduction

India is home to an extraordinary diversity of languages, including more than 460 tribal languages documented in the 2011 Census. Many of these languages have substantial speaker populations, yet they remain severely underrepresented in modern NLP research. The primary reasons include limited digital presence, inconsistent or evolving orthographies, and the absence of large, standardized parallel corpora (Joshi et al., 2020; Nekoto et al., 2020). As a result, the rapid progress in deep learning and multilingual LLMs has had little impact on the communities that speak these languages.

The challenges faced by tribal communities extend beyond language processing. Data scarcity affects crucial sectors such as healthcare, education, biodiversity monitoring, and governance. Practitioners working in these areas often operate in environments with limited connectivity and minimal technical infrastructure. For AI systems to be useful in such contexts, they must be resilient to missing modalities, noisy inputs, and shifting distributions (Sun et al., 2023). This shared task therefore focuses not only on translation accuracy but also on robustness under real-world constraints.

The MMLoSo Language Challenge 2025 (Shukla et al., 2025) aims to directly address this gap by advancing research in machine translation for India's tribal and very low-resource languages. While recent large-scale multilingual and multimodal models (Conneau et al., 2020; Alayrac et al., 2022; Li et al., 2023) have demonstrated impressive generalization abilities, they are primarily trained on high-resource languages and domains. Tribal languages, which are often oral, regionally grounded, and culturally specific, fall outside the distribution of mainstream training datasets. This disconnect results in poor performance, hallucination, and limited utility for real-world applications.

## 2 Task Description

This shared task focuses on developing neural machine translation systems for bidirectional translation between high-resource languages (Hindi, English) and four low-resource Indian tribal languages: Bhili, Mundari, Gondi, and Santali. The primary challenge lies in building effective models despite limited parallel data, diverse writing

*Proceedings of MMLoSo 2025: First Workshop on Multimodal Models for Low-Resource Contexts and Social Impact*, pages 121–129
December 23, 2025 ©2025 Association for Computational Linguistics

systems, and domain-specific vocabulary characteristic of tribal communities.

The specific goals of the shared task are:

1. **Develop effective translation systems for low-resource Indian languages.** Models should handle small corpora, domain variation, and diverse scripts while producing faithful and stable translations.

2. **Evaluate cross-lingual transfer and parameter-efficient adaptation.** This includes approaches such as LoRA-based fine-tuning, multilingual joint training, retrieval augmentation, and hybrid SMT-NMT pipelines.

3. **Promote socially grounded and inclusive NLP.** Better translation systems can support information access in healthcare, education, disaster response, and public services, helping reduce the digital divide faced by tribal communities.

4. **Create open foundations for future research.** By releasing curated datasets and baseline systems, the shared task aims to enable long-term work in tribal language processing and cultural preservation.

### 2.1 Task Format

Participants were provided with four training files (one per language pair: Bhili–Hindi, Mundari–Hindi, Gondi–Hindi, and Santali–English), each containing parallel sentences with columns: **row_id** (unique identifier), a high-resource language column (`hindi` or `english`), and the corresponding tribal language column.

The test set provided unlabeled source sentences with columns: **row_id**, **source_sentence**, **source_lang**, and **target_lang**. Systems were expected to generate translations in both directions (high-to-low and low-to-high resource) and submit predictions with the format: **row_id**, **source_lang**, **source_sentence**, **target_lang**, and **target_sentence**.

### 2.2 Evaluation Metrics

Systems were ranked using a weighted composite score that balances word-level accuracy and character-level fluency. The final score is defined as:

| **Lang Pair** | **Columns** |
|---|---|
| Hin ↔ Bhili | row_id, hindi, bhili |
| Hin ↔ Mundari | row_id, hindi, mundari |
| Hin ↔ Gondi | row_id, hindi, gondi |
| Eng ↔ Santali | row_id, english, santali |
| Test Data | row_id, source_sent |

Table 1: Dataset Overview.

$$S = 0.6 \times \text{BLEU} + 0.4 \times \text{chrF}$$

BLEU (Papineni et al., 2002) receives higher weight (60%) due to its sensitivity to n-gram precision, while chrF (Popovic, 2015) contributes 40% to capture character-level morphology—crucial for languages like Mundari and Santali.

Scores were computed for each translation direction and aggregated with an additional directional weighting:

- **High-to-Low (60%):** Higher weight is assigned to translations *into* the tribal languages, reflecting the task's emphasis on accessibility for low-resource communities.

- **Low-to-High (40%):** Translations *from* the tribal languages are given slightly lower weight.

This composite metric ensures that systems are evaluated fairly across both lexical accuracy and morphological sensitivity.

Full Formula:

$$\text{Score} = 0.6\left[0.6\sum_{d\in H\to L}\text{BLEU}_d + 0.4\sum_{d\in L\to H}\text{BLEU}_d\right] + 0.4\left[0.6\sum_{d\in H\to L}\text{chrF}_d + 0.4\sum_{d\in L\to H}\text{chrF}_d\right].$$

Where $H \to L$ represents the high-to-low directions (e.g., Hindi → Bhili) and $L \to H$ represents the low-to-high directions.

## 3 Dataset

The shared task focuses on four low-resource tribal languages of India (see Table 1): Bhili, Mundari, Gondi, and Santali. All four have large speaker communities but extremely limited digital presence. Existing text is sparse, unaligned, or not machine readable, making these languages particularly challenging for machine translation. The dataset comprises 80,000 training pairs (20,000 per language pair) and 15,999 test sentences distributed across bidirectional translation tasks.

These language pairs represent a mix of typologically related and unrelated languages. Hindi, Bhili, and Mundari are written in Devanagari, though they belong to different language families (Indo-Aryan and Austroasiatic). Gondi, also written in Devanagari, is Dravidian. Santali, however, is an Austroasiatic language written in the Ol Chiki script and paired with English, making it one of the most linguistically distant settings in the task.

We provide a brief overview of the languages used in this challenge below:

***Bhili (ISO 639-3: bhb)*** is a Western Indo-Aryan language spoken by approximately 13 million people across western and central India. Despite its large speaker base and proximity to major languages like Gujarati and Marathi, Bhili remains severely under-resourced in terms of digital content and NLP tools. The variety used in this dataset is the Jhabua dialect from Madhya Pradesh, written in Devanagari script.

***Mundari (ISO 639-3: unr)*** is a North Munda language of the Austroasiatic family, spoken by approximately 1.6 million people primarily in Jharkhand and neighboring states. Although traditionally written in multiple scripts, we use Devanagari to maintain consistency with Hindi-aligned NLP pipelines. Mundari has complex word structures where multiple prefixes and suffixes combine to form single words, and suffers from limited parallel data and strong influence from regional dominant languages.

***Gondi (ISO 639-3: gon)*** is a South-Central Dravidian language with approximately 3 million speakers distributed across central India. Due to language shift pressures, many speakers are shifting to regional dominant languages. We focus on the Devanagari-script variety to ensure compatibility with mainstream Hindi NLP tools and resources.

***Santali (ISO 639-3: sat)*** is a major Munda language with over 7 million speakers across India, Bangladesh, Nepal, and Bhutan. Unlike the other three languages in this task, Santali is paired with English rather than Hindi. We focus on the Ol Chiki script, an indigenous writing system created in the 1920s that is increasingly used in education and modern publications. NLP resources for Ol Chiki remain extremely limited, particularly for parallel corpora.

### 3.1 Data Collection and Preprocessing

All parallel datasets were sourced from curated web content, Wikipedia articles, and resources provided by the Ministry of Tribal Affairs, Government of India. Each dataset follows a uniform structure with a unique row ID and a pair of parallel sentences. The data underwent a multi-stage preprocessing pipeline: sentence alignment, character normalization, deduplication of near-identical pairs, and filtering to retain only sentences between 6 and 80 words. Cosine similarity filtering removed pairs that were overly similar across source and target. For Bhili, expert-driven translation was employed to ensure accuracy and contextual quality. Tokenization was handled uniformly using SentencePiece (Kudo and Richardson, 2018).

For the English–Santali pair, the English source was produced by translating curated Hindi content using IndicTrans2 (Gala et al., 2023) to maintain thematic consistency across language pairs, with additional lowercasing applied to English text. The test set contains only the source sentence and language direction; participants must generate the target translation. It was created using stratified sampling across multiple domains to ensure coverage while avoiding repetition. All data is distributed under the Creative Commons BY-SA 4.0 license.

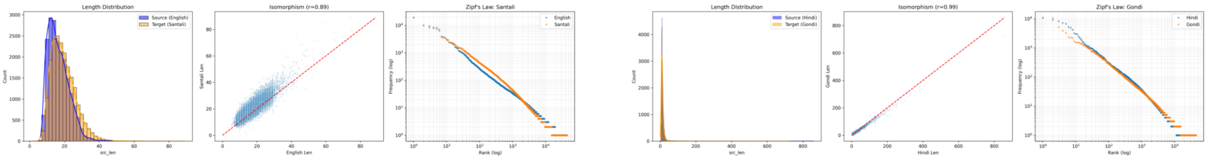
### 3.2 Dataset Statistics

Tables 2 and 3 summarize the corpus statistics. All training sets contain exactly 20,000 parallel pairs. Vocabulary analysis (Figure 1) reveals significant lexical sparsity, particularly in Bhili and Mundari, where target vocabularies exceed 60,000 tokens due to high morphological variation. Santali also exhibits high diversity driven by productive affixation in the Ol Chiki script. Structurally, Gondi remains the most compact (avg. 13.8 tokens), whereas Bhili presents the longest sequences in both source and target directions.

Overall, the dataset reflects real-world low-resource conditions: diverse scripts, high morphological richness, and large vocabularies relative to corpus size. These properties justify the need for specialized training strategies such as LoRA-based direction-specific fine-tuning, retrieval-augmented methods, and conservative decoding techniques.

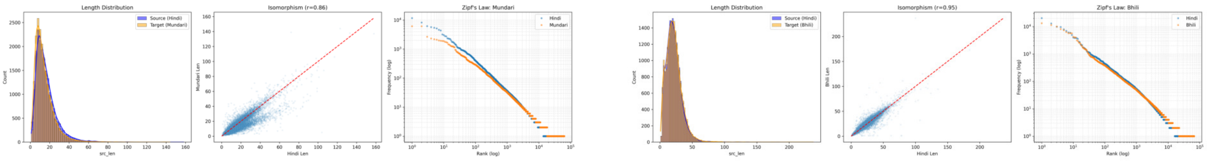
### 3.3 Dataset Complexity Analysis

We present a comprehensive exploratory data analysis (EDA) of the results to understand the linguistic



English–Santali

Hindi–Gondi



Hindi–Mundari

Hindi–Bhili

Figure 1: EDA visualizations across the four language pairs.

| Dataset | Src | Src Len | Src Voc | Tgt | Tgt Voc |
|---|---|---|---|---|---|
| santali-train | Eng | 16.5 | 42114 | San | 44764 |
| gondi-train | Hin | 14.4 | 25133 | Gon | 44756 |
| mundari-train | Hin | 16.3 | 35373 | Mun | 63426 |
| bhili-train | Hin | 21.3 | 40636 | Bhi | 67019 |

Table 2: Training data statistics (20,000 pairs per language pair).

| Source Lang | Count | Avg Len | Vocab Size |
|---|---|---|---|
| Bhili | 1999 | 21.6 | 13125 |
| English | 2000 | 15.5 | 7810 |
| Gondi | 2000 | 6.9 | 5673 |
| Hindi | 6000 | 15.7 | 16860 |
| Mundari | 2000 | 14.1 | 11356 |
| Santali | 2000 | 14.4 | 7247 |

Table 3: Test set statistics

complexity of the corpus. Table 4 summarizes the metrics calculated using the NLLB tokenizer and standard lexical diversity measures.

### Morphological Complexity

We computed *Token Fertility* - the average number of subword tokens required to represent a single word—as a proxy for morphological richness. As shown in Table 4, Santali and Mundari show extremely high fertility rates (3.02 and 2.55, respectively) compared to their source languages. This indicates highly agglutinative structures where a single word often translates to multiple tokens in the NLLB vocabulary, posing a significant challenge for the model.

### Structural Divergence

We measured the Pearson correlation coefficient ($r$) between source and target sentence lengths.

- **High Isomorphism** ($r > 0.9$)**:** Hindi-Gondi ($r = 0.99$) and Hindi-Bhili ($r = 0.95$) show near-perfect length correlation. This suggests these languages share similar syntactic structures with Hindi, making them easier for models to align.
- **Divergence** ($r < 0.9$)**:** English-Santali ($r = 0.89$) and Hindi-Mundari ($r = 0.86$) show lower correlation, reflecting significant structural differences (e.g., SVO vs. SOV word order) that complicate translation.

Table 4: Linguistic Statistics of the Training Data. **TTR**: Type-Token Ratio (Lexical Richness). **Fertility**: Avg. subwords per word (NLLB tokenizer).

| **Pair** | **Corr** ($r$) | **TTR** | | **Fertility** | |
|---|---|---|---|---|---|
| | | Src | Tgt | Src | Tgt |
| Eng-Santali | 0.89 | 0.127 | 0.116 | 1.35 | **3.02** |
| Hin-Gondi | **0.99** | 0.088 | 0.162 | 1.39 | 2.32 |
| Hin-Mundari | 0.86 | 0.108 | **0.223** | 1.50 | 2.55 |
| Hin-Bhili | 0.95 | 0.095 | 0.155 | 1.43 | 1.73 |

## 4 Approaches and Results

The shared task attracted 18 teams who submitted solutions on Kaggle. Teams achieving a private leaderboard score above 150 were invited to submit challenge papers describing their approaches. Out of the seven eligible teams, four accepted this invitation and provided detailed system descriptions. Most teams started from multilingual pretrained models and fine-tuned them on the provided training data, adding specialized techniques to handle data scarcity, morphological complexity, and vocabulary mismatch. Table 5 presents the final rankings for the four teams that submitted papers.

Table 5: Complete final leaderboard ranked by private scores. Teams marked with * submitted challenge papers (invitation threshold: private score >150).

| Rank | Team Name | Public | Private |
|---|---|---|---|
| 1 | SajayR* | 319.39 | **212.04** |
| 2 | HCMUS_PrompterX* | 311.61 | 186.37 |
| 3 | No | 310.19 | 184.87 |
| 4 | boy Magic | 309.98 | 179.83 |
| 5 | Shooting star* | 216.04 | 179.49 |
| 6 | VaibhavKanojia* | 171.39 | 161.11 |
| 7 | Shivansh Jha | 193.36 | 153.90 |
| 8 | Badr A | 224.09 | 145.47 |
| 9 | EROL | 155.22 | 134.02 |
| 10 | king | 162.25 | 133.96 |
| 11 | e0nia | 143.89 | 131.56 |
| 12 | c00k | 290.30 | 111.47 |
| 13 | Kabir Raj Singh | 140.64 | 104.07 |
| 14 | Daglox Kankwanda | 122.63 | 96.22 |
| 15 | Michael Ibrahim | 92.00 | 87.17 |
| 16 | Harsh Rajbhar | 37.28 | 35.18 |
| 17 | winner_can_exist | 26.10 | 20.80 |
| 18 | Code by Nadiia | 4.22 | 6.28 |

## 4.1 System 1: LoRAs in All Directions

The winning team (Table 5, Rank 1) started from the widely used NLLB (Costa-jussà et al., 2022) backbone but later switched to the IndicTrans2 model (1.1B parameter version) (Gala et al., 2023). This decision was driven by a detailed analysis of tokenization behavior on the provided datasets. IndicTrans2 showed far superior fertility statistics, especially for Santali. For example, the model produced an average fertility of 1.44 tokens per character for Santali, compared to 3.07 for NLLB. Lower fertility indicates that the model represents the script more efficiently and reduces fragmentation, which is essential when dealing with agglutinative morphology. This gave IndicTrans2 a clear representational advantage.

### 4.1.1 Training Strategy

The team employed a three-stage "Saturate-then-Specialize" pipeline designed to balance multilingual generalization with task-specific specialization.

**Tag-Only Preprocessing.** To preserve contrastive linguistic properties often lost during script unification, the team utilized explicit language tags. Surrogate tags were mapped to the nearest vocabulary equivalent (e.g., Bhili to `mar_Deva`), enabling the model to internally align unseen languages with typologically similar ones without external transliteration.

**Joint Fine-Tuning.** The backbone was first fine-tuned on the union of all datasets to “saturate” the model with shared domain knowledge. This exposure to diverse lexical distributions stabilized the embedding space, improving robustness in data-sparse conditions.

**Directional LoRA Adapters.** To eliminate catastrophic forgetting, the backbone was frozen while training separate Low-Rank Adapters (LoRA) (Hu et al., 2022) for each direction ($r = 64, \alpha = 128$). This parameter isolation prevented cross-lingual interference (negative transfer) while allowing targeted adaptation for specific scripts and grammars.

### 4.1.2 Inference Strategy

To mitigate hallucinations, the system employed noisy channel reranking (Yee et al., 2019). Instead of relying solely on the forward probability $P(y|x)$, $K$ candidates were generated and re-scored using a reverse model $P(x|y)$:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} [\alpha \cdot \log P(y|x) + \beta \cdot \log P(x|y)]$$

This objective acts as a semantic regularizer by rewarding translations from which the source sentence is reconstructible. Candidates that hallucinate or deviate in meaning yield low reverse probabilities and are effectively filtered out.

## 4.2 System 2: JHARNA MT

The runner-up team (Rank 2) identified hallucination as the primary failure mode in data-scarce settings and designed a hybrid architecture. By integrating retrieval-based memory and Statistical Machine Translation (SMT) with modern Neural Machine Translation (NMT), the system grounds neural generations in real training examples to ensure lexical fidelity.

### 4.2.1 Training Strategy

The pipeline prioritizes observed data via a two-tier retrieval module:

- **Exact Match:** Direct retrieval of training targets for inputs seen during training ( 8% of test data).
- **Fuzzy Match:** Retrieval of examples with edit distance $\leq 1$ to ground common expressions.

For unseen inputs, candidates are generated by two distinct models:

- **SMT Branch:** An IBM Model 1 system (Brown et al., 1993) with diagonal alignment priors. This branch produces "literal" translations that are robust against NMT hallucinations.
- **NMT Branch:** An NLLB-200 (Costa-jussà et al., 2022) model fine-tuned with LoRA adapters (Hu et al., 2022) ($r = 16$) to provide fluency and context awareness.

#### 4.2.2 Inference Strategy

To unify the branches, the system pools $N$ hypotheses and applies Minimum Bayes-Risk (MBR) decoding (Kumar and Byrne, 2004). This selects the "consensus" translation that maximizes average similarity (BLEU/chrF) to all other candidates. This effectively filters out hallucinations, as they tend to diverge significantly from the stable SMT outputs.

The team observed complementary error modes: SMT is rigid but lexically safe, while NMT is fluent but prone to hallucination. The hybrid MBR approach successfully combines these strengths, yielding significant stability improvements over pure neural baselines.

### 4.3 System 3: Breaking Language Barriers

The Breaking Language Barriers team (Rank 5) adopted a data-centric strategy using NLLB-200 (Costa-jussà et al., 2022) as the primary model and mBART-50 (Liu et al., 2020) as a validator. Lacking native proficiency, the focus remained on detecting pathological model behaviors through statistical heuristics rather than linguistic intuition.

#### 4.3.1 Training Strategy

To address data sparsity, all bidirectional pairs were concatenated into a single $D_{\text{unified}}$ dataset. This doubled the effective training size and enforced a shared semantic space across all languages, encouraging cross-lingual alignment and reducing overfitting.

#### 4.3.2 Inference Strategy

A “Safety Net” ensemble was implemented to filter catastrophic failures. If the primary model’s output length ratio was extreme ($< 0.3$ implying under-generation or $> 3.0$ implying hallucination), the system fell back to mBART predictions. This mechanism significantly improved robustness, boosting the private leaderboard score by over 5 points.

### 4.4 System 4: Divide and Translate

The Divide and Translate team (Rank 6) based their system on the hypothesis that multilingual models suffer from negative interference when dealing with languages that differ significantly in syntax. For example, English is an SVO language while Santali and the Hindi aligned languages are SOV. To avoid cross language interference, the team trained *separate decoders* for each direction while keeping a shared encoder.

#### 4.4.1 Training Strategy

The shared encoder from NLLB-600M (Costa-jussà et al., 2022) was frozen completely. This prevented the degradation of the multilingual representations learned during large-scale pretraining. Only the decoder parameters were updated. This acted as a strong regularizer and reduced the risk of overfitting, which is common when training on small datasets.

#### 4.4.2 Results

The approach produced very stable results with minimal public private score variance. However, the performance ceiling was lower than that of LoRA based or hybrid systems. The frozen encoder limited the model’s ability to adapt to the specific scripts and morphological patterns of the tribal languages.

## 5 Analysis and Key Findings

This section examines common patterns, challenges, and strategic decisions across the four submitted systems. We analyze the key factors that influenced performance, including choice of backbone models, approaches to mitigating hallucination, and how teams addressed script diversity and morphological complexity. These insights highlight both successful strategies and persistent challenges in low-resource machine translation.

### 5.1 Model Architecture Choices

A major distinction among the systems was the choice of backbone model. Three teams used NLLB (Costa-jussà et al., 2022), but the winning system built on IndicTrans2 (Gala et al., 2023). The differences stem from tokenization behavior. Many teams reported that NLLB struggled to tokenize Ol

Chiki and certain Devanagari variants. IndicTrans2, pretrained on a broader set of Indic languages, handled these scripts much more effectively.

Submissions also followed two main architectural strategies: **(1) Pure neural systems**, which rely solely on model scale, architecture, and decoding techniques; and **(2) Hybrid systems**, which combine SMT, retrieval, and neural models to address data scarcity. Hybrid approaches showed strong performance in low-resource settings, demonstrating that classical MT components remain useful when training data is limited.

### 5.2 Hallucination Mitigation Strategies

Hallucination was the dominant challenge of the task. Each system proposed a different mitigation strategy:

- **System 1:** Reverse model scoring (Noisy Channel).
- **System 2:** Consensus decoding (MBR).
- **System 3:** Length based fallbacks (Conservative Ensemble).
- **System 4:** Architectural constraints (Frozen Encoder).

The most effective strategies were the Noisy Channel and MBR approaches, both of which rely on generating multiple candidates and filtering them via external validation signals.

### 5.3 Script and Morphological Challenges

Santali, written in Ol Chiki, presented a major challenge for many backbones. Models without explicit support for the script required additional preprocessing or surrogate tags. IndicTrans2 had a clear advantage due to its broader coverage of Indic scripts.

Mundari has a very high type token ratio, more than double that of Hindi. This causes data sparsity and increases the difficulty of learning accurate word representations. System 2 used back translation to expand the dataset and expose the model to a wider range of morphological patterns.

### 5.4 Generalization and Overfitting

Some teams observed significant score drops between the public and private leaderboards. This indicates overfitting to specific patterns in the public test set. The Divide and Translate system showed the smallest gap, supporting their claim that encoder freezing works as a strong regularizer. The large gap between Public and Private scores (Table 5) highlights the challenge of generalization in low-resource settings.

## 6 Conclusion

The MMLoSo 2025 Shared Task is one of the first efforts to evaluate machine translation for India's tribal and severely low-resource languages. By releasing parallel datasets for Bhili, Gondi, Mundari, and Santali, the task provides a consistent benchmark for languages rarely covered in mainstream NLP. The task attracted 18 participating teams, with the top four achieving private scores above 150 and contributing detailed system descriptions. Teams employed low-resource-focused techniques such as retrieval augmentation, hybrid SMT–NMT systems, noisy-channel reranking, and direction-specific LoRA adapters to address challenges including script diversity, limited digital text, and rich morphology.

The results highlight the need for more documentation and larger datasets for tribal languages, which have large speaker communities but limited digital resources. By offering open data, standardized evaluation, and strong baselines, the shared task aims to support long-term research in inclusive and socially meaningful NLP. Future editions will expand toward multimodal translation, cross-dialect evaluation, and domain-specific benchmarks for governance, education, and cultural preservation.

## 7 Acknowledgements

We thank the Ministry of Tribal Affairs, Government of India, and especially Shri Vibhu Nayar for their guidance and support. We also acknowledge the Tribal Research Institutes of Odisha, Jharkhand, Chhattisgarh, and Madhya Pradesh, along with all translators and participants, for their essential contributions to this work.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand

Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Jay P. Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M., Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Trans. Mach. Learn. Res.*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Shankar Kumar and William J. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2004, Boston, Massachusetts, USA, May 2-7, 2004*, pages 169–176. The Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi E. Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Z. Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan Van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing K. Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2144–2160. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Ankita Shukla, Sandeep Kumar, Amrit Singh Bedi, and Tanmoy Chakraborty. 2025. Multimodal models for low-resource contexts and social impact 2025 machine translation challenge. Kaggle Competition. MMLoSo Workshop Co-located with IJCNLP-AACL 2025.

Lixu Sun, Nurmemet Yolwas, and Lina Jiang. 2023. A method improves speech recognition with contrastive learning in low-resource languages. *Applied Sciences*, 13(8):4836.

Kyra Yee, Yann N. Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. 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, Hong Kong, China, November 3-7, 2019*, pages 5695–5700. Association for Computational Linguistics.

# Author Index

Ahmed, Kawsar, 35
Ahsan, Shawly, 35
Al Sefat, Abdullah, 58
Alabi, Jesujoba Oluwadara, 25
Anikina, Tatiana, 25

Baghel, Shruti Singh, 86
Bedi, Amrit Singh, 121
Bhavsar, Arnav, 86

Chakraborty, Tanmoy, 121
Chatterjee, Sandeep, 121
Cheema, Gullal S., 121
Chikkala, Ravikiran, 25

Dinh, Dien, 114
Dương, Nguyễn Đình Hà, 114

Ekbal, Asif, 95

Goyal, Pawan, 86

Hoque, Mohammed Moshiul, 35
Huynh, Trung Kiet, 114

Jacob, Jubal Chandy, 48
Jena, Sushovan, 86
Jennings, Nicholas, 25

Kamau, Paul, 106
Kamigaito, Hidetaka, 13
Kanojia, Vaibhav, 109
Kumar, Sandeep, 121
Kurian, Betty, 78

Lam, Phu Quy Nguyen, 114

Minh, Dao Sy Duy, 114
Mohiuddin, Md., 35

Nakai, Toshiki, 25
Nguyen, Long HB, 114
Nguyen, Tran Chi, 114

Oberkircher, Lena, 25

Padó, Sebastian, 1
Pakray, Partha, 95
Pham, Phu-Hoa, 114
Pradhan, Anurag, 86

Raj, Sajay, 101
Rathore, Yash Pratap Singh, 86

Sakajo, Haruki, 13
Sengupta, Abhijeet, 78
Shukla, Amit, 86
Shukla, Ankita, 121
Singh, Kshetrimayum Boynao, 95
Singh, Pooja, 121
Skachkova, Natalia, 25
Soda, Komei, 13

Takato, Hiroshi, 13
Tang, Ming Ze, 48
Thomas, Kerstin, 1
Tsutsui, Hiroshi, 13

Upadhyay, Abhinav, 78

Watanabe, Taro, 13