ANLP Assignment 4 Report: Multi-lingual and Multi-cultural Figurative Language Understanding

Anvesha Katariyar
aakatari@andrew.cmu.edu

Prachiti Garge
pgarge@cs.cmu.edu

Shreya Terupally
sterupal@cs.cmu.edu

1 Introduction

Figurative language, integral to human communication, remains underexplored in NLP, especially across diverse linguistic and cultural contexts. Multilingual figurative question answering (figQA) presents a critical challenge due to the intricacies of cross-cultural interpretation and contextual understanding. This project aims to address the lacuna in research by investigating multilingual figurative QA.

Through a literature survey, we delve into existing methodologies and challenges in this domain. We leverage the Multilingual FigQA dataset to reproduce competitive baselines and propose three novel techniques to address the limitations. 1) Contrastive finetuning, which leverages contrastive learning to better capture figurative semantics; 2) Chain-of-Thought (CoT) prompting, which encourages models to reason step-by-step, aiding in contextual understanding; and 3) Document augmentation, which supplements the input with relevant cultural and linguistic information.

Detailed analysis is performed to understand the strengths and weaknesses of each technique. By advancing multilingual figQA capabilities, our work contributes to developing more culturally-aware and context-sensitive NLP models.

2 Previous work

2.1 Figurative Language Understanding

(Aghazadeh et al., 2022) delves into the rich metaphorical nature of human languages and how such expressions facilitate understanding by linking novel concepts to more familiar ones. It explores the hypothesis that large pre-trained language models (PLMs) harbor metaphorical knowledge beneficial for natural language processing (NLP) systems. The investigation spans multiple datasets dedicated to metaphor detection and encompasses four diverse languages: English, Spanish, Russian, and Farsi.

In tackling the nuances of figurative language, FLUTE (Chakrabarty et al., 2022) presents a dataset designed to probe the true comprehension of figurative expressions in NLP. Addressing the limitations of existing benchmarks plagued by spurious correlations, FLUTE introduces 9,000 instances of figurative natural language inference across four categories: Sarcasm, Simile, Metaphor, and Idioms. Unique for including textual explanations, the dataset, created through a model-in-the-loop framework leveraging GPT-3, crowdworkers, and expert annotators, ensures models grasp the intended meanings for the right reasons.

BirdQA (Zhang and Wan, 2022) further enriches the landscape with a bilingual dataset targeting the understanding of riddles by QA models. Encompassing 6,614 English and 8,751 Chinese riddles paired with distractors from Wikipedia, BiRdQA challenges models with the figurative language and creative complexity of riddles, featuring misleading elements like metaphors and puns that demand advanced reasoning. Despite minimal human intervention in its creation, initial tests reveal a significant gap between human and machine performance, highlighting the dataset's potential to advance NLP. By delving into the intricate understanding of riddles, BiRdQA offers a distinctive resource to the QA and NLP communities, fostering the development of models with enhanced comprehension and reasoning capabilities across languages.

2.2 Multilingual NLP systems

The field of multilingual natural language processing (NLP) employs various strategies and frameworks to tackle the complexities introduced by linguistic diversity and unequal resource availability. Notable models in this domain include mBERT (Multilingual BERT) (Devlin et al., 2018), an extension of the original BERT , trained on texts from 104 languages for multilingual understanding, and XLM-Roberta (XLM-R) (Conneau et al., 2019), an extension of the RoBERTA architecture, leveraging a vast corpus from over 100 languages and displaying significant performance gains for cross-lingual transfer tasks. We evaluate our task performance on 2 variants (base and large) of this model.

Additionally, GPT-3 (Brown et al., 2020), a large-scale causal decoder-only model with 175B parameters, demonstrates some multilingual capabilities due to the inclusion of non-English content in its training dataset, while its successor, GPT-3.5, refines these capabilities with architectural improvements and an updated training corpus. XGLM (Lin et al., 2022), a multilingual generative language model purposefully trained on a diverse multilingual corpus covering 30 languages, outperforms GPT-3 of comparable size in multilingual commonsense reasoning and natural language inference. Moreover, BLOOM (BigScience et al., 2023), with its vast scale of 176 billion parameters and support for 46 natural languages and 13 programming languages, is dedicated to fostering inclusivity and open access, although its immense size presents challenges for comprehensive evaluation.

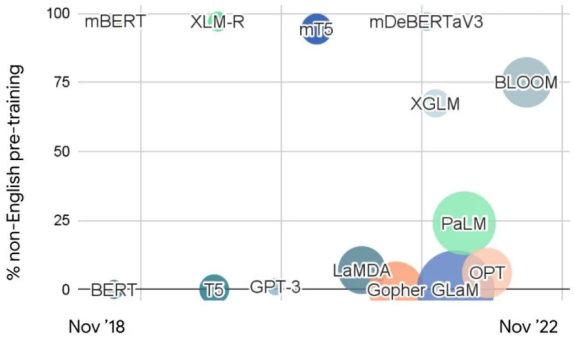

Figure 1: Multilingual Large Language Model

2.3 Multilingual Figurative QA Benchmark

MABL (Kabra et al., 2023) is a recent a figurative language inference dataset, for seven diverse languages associated with a variety of cultures: Hindi, Indonesian, Javanese, Kannada, Sundanese, Swahili and Yoruba. This dataset focuses on cultural and regional concepts for figurative expressions, and formulates the task as QA task with a choice between 2 possible meanings for a metaphor. The dataset also performs analysis that which languages' concept sets are similar to others, and finds that general highest overlapping languages originating from the same region.

We use this benchmark to evaluate our models and understand integration of cultural knowledge into today's multlingual LLMs.

	Figurative Expression	Inference
yo	Omah iku kaya istana (The house is like a palace.)	Omah iku apik banget. (The house is very nice.) Omah iku elek banget. (The house is very ugly.)
id	Rambutnya seperti bihun. (Her hair is like vermicelli.)	Rambutnya keriting. (Her hair is curly.) Rambutnya lurus. (Her hair is straight.)
hi	जीवन मीठा गुलकन्द है। (Life is sweet Gulkand.)	जीवन अच्छा है। (Life is good.) जीवन बुरा है। (Life is bad.)
kn	ಅದು ದೋಸೆಯಂತೆ ಗರಿಗರಿಯಾಗಿತ್ತು. (It was crispy like a dosa.)	ಅದು ಗರಿಗರಿಯಾಗಿದೆ (It is crisp.) ಅದು ಗರಿಗರಿಯಾಗಿರಲಿಲ್ಲ (It was not crisp.)
sw	Maneno yake ni sumu. (His words are like poison.)	Maneno yake yanaponya. (His words heal.) Maneno yake yanaangamiza. (His words are devastating.)

Figure 2: Example of figurative expressions in Multilingual FigQA benchmark

3 Baselines Analysis

3.1 Evaluation Methodology

We evaluate 5 baseline multilingual large language models (LLMs) on our MABL benchmark. Previous literature has primarily concentrated on the performance of encoder-based and GPT models, while this project broadens the evaluation to include recent LLMs such as XGLM. The experiments conducted aim to assess the performance of these models on a multilingual dataset. The evaluation considers zero-shot performance, where

encoder-based models are finetuned on English labeled data and processed in a specific format, while decoder-based models concatenate the metaphor and potential meanings, computing log probabilities for prediction.

Additionally, few-shot evaluation is explored for the best-performing and least-performing models for both encoder-based and decoder-based categories. Few-shot evaluation for encoder-based models is performed by finetuning with few examples in the target language alongside English data, and utilizing in-context learning for decoder-based models. The experiments provide insights into the capabilities of these multilingual LLMs on the multilingual benchmark, expanding upon previous research focused primarily on encoder-based and GPT models.

3.2 Baseline Results and Error Analysis

We summarize the results of our baseline models in the Appendix Tables.

Performance generally increases with model size . The zero-shot performance evaluation reveals that different languages have varying best-performing models, primarily tied between GPT-3.5 and XLM-R-large, indicating that performance generally increases with model size but is greatly impacted by differences in training data and model architectures. **GPT-3.5 excels on English, Indonesian, and Swahili**, likely due to their conceptual similarity to English and GPT-3.5's highly parameter-tuned nature. Conversely, **XLM-R-large outperforms GPT-3.5 on Hindi, Kannada, and Sundanese**, suggesting that GPT-3.5 is not a universally effective multilingual model and that cultural concept shifts between English and these languages may contribute to the performance differences. For the low-resource language Yoruba, all models perform close to random. Surprisingly, XGLM-large, purposefully trained on a multilingual corpus, underperforms compared to GPT-3.5 for non-English languages, potentially due to GPT-3.5's closed-source nature and multilingual-specific performance improvements. Additionally, hyperparameter choices may impact XGLM-large's performance.

Few-shot performance In the few-shot performance evaluation, models generally showed little improvement with additional examples. While GPT-3.5 demonstrated some gains for certain languages through in-context learning, the final accuracy remained close to random for all models, suggesting limitations in incorporating multicultural knowledge through few-shot finetuning.

Qualitative Analysis XLM-R showcases good performance in English and Indonesian, and decent performance across all languages. **XLM-R fails on medium and hard examples** due to lack of semantic and figurative knowledge understanding, especially for low-resource languages like Sundanese. **GPT 3.5 struggles to understand semantic meaning in easy examples for Indic low-resource languages** like Kannada and Hindi, but it performs well on languages with higher conceptual similarity to English. XGLM-large, likely due to its balanced and diverse training corpus, performs moderately across all languages but fails on cases requiring culture-specific interpretations, language comprehension issues, and cultural knowledge. For XLM-R, finetuning on culturally rich text and using intermediary similar languages could help, while chain-of-thought reasoning and in-context learning might benefit GPT-3.5 and XGLM for low-resource languages. The final accuracy remains close to random for mBERT, indicating limitations in incorporating multicultural knowledge.

4 Our Proposed Model

4.1 Contrastive Finetuning for XLM-R models

We observe that the current baseline models do not tackle the embedding space in encoder-only models, but rather modify the downstream task-specific module for both zero-shot and few-shot learning. Ideally, we need to ensure that the embeddings of both metaphor and meaning are closely aligned in the embedding space. To address this, we propose to finetune the base model layers using a contrastive loss, trying to reduce the distance between similar meaning and metaphor embedding. We conduct an ablation-study to determine the percent of Multlingual FigQA

Model	Input	Improved Prediction	Original Prediction	Reasoning
Finetuned XLM-R-base (su)	Omongan lalaki téh ibarat sarang madu	Omongan lalaki téh ibarat sarang madu	Omongan lalaki téh ibarat sarang madu	Similar concepts across training data
Finetuned XLM-R-large (id)	Tugas mahasiswa bagaikan bisa dihitung jari	Tugas mahasiswa sedikit	Tugas mahasiswa banyak	Improved cultural understanding
Finetuned XLM-R-large (en)	The teacher is as encouraging as a dead bird	the teacher doesn't help	the teacher helps	Simple metaphor
Finetuned XLM-R-large (jv)	Banyune kolam renang kaya banyu leri	Banyune kolam renang bening	Banyune kolam renang buthek	Increased semantic understanding
Cultural Document Augmentation (su)	Duit bisa ngarobah jalma kawas Qarun anu dikubur ku hartana sorangan	Duit bisa ngarobah jalma jadi sarakah	Duit bisa ngarobah jalma jadi bersukur	Better cultural references understanding
Chain-of-thought	उसका गुस्सा बर्फ की सिल्ली सा है।	उसका गुस्सा जल्दी उतर जाता है।	उसको गुस्सा बहुत ज्यादा आता है।	Established logical connection between metaphor and option

Table 1: Examples where our proposed method improved over its predecessor.

dataset to use as the training material for our task.

4.2 In-context learning for Decoder-only Models

We observe that direct few-shot in-context learning does not effectively address the issue of conceptual shift. However, we hypothesize that the model possesses the capacity for in-context learning across multiple languages, but the context provided is not sufficiently comprehensive. To examine this, we plan to enhance the context with the following strategies:

- Employing few-shot and zero-shot chain-of-thought prompting and incorporate cultural reasoning for each example.
- Generating a hypothetical (or curated) document per language that catalogues culture-specific objects (eg. Neem in Hindi) and associations, and provide as context. This document can be manually crafted, produced by an LLM (similar to Self-Instruct), or retrieved online.

5 Results

Table 2 and Table 3 display the results of our proposed methods on different languages. Table 1 contains few successful examples where our proposed method improved over the predecessor method. We also analyze the error cases of our proposed method in Table 4.

5.1 Ablation Study of Contrastive Finetuning

We conducted an ablation study to investigate the impact of training data size on the performance of contrastive fine-tuning using the XLM-R base model. The dataset comprised instances from multiple languages, with the aim of enhancing cross-lingual transfer learning. We evaluated the model's performance on held-out test data comprising 50% of the instances, while varying the training data size from 5% to 50% of the total data. Our findings revealed that as the training data size increased, the model's performance generally improved. However, this improvement was not uniform across all languages. Languages

	XLM-R-base	Fine-tuned XLM-R-base	XLM-R-large	Fine-tuned XLM-R-large
en_dev	71.13	74.6	78.05	82.16
hi	58.2	58.2	67.0	65.1
id	69.3	75.23	76.4	81.6
jv	52.1	60.16	57.33	64.72
kn	57.25	59.0	60.3	59.5
su	54.45	61.7	52.82	64.1
sw	51.58	54.6	58.41	63.5
yo	-	-	-	-

Table 2: Performance of contrastive finetuned XLM-R models on different languages. Training and testing data was chosen to be a 30:70 split. The numbers are on the test subset of the MABL dataset.

with a similar script and concept set as English, such as Indonesian, Javanese, and Sundanese, exhibited the most significant performance gains from contrastive fine-tuning. In contrast, we observed a decline in performance for Hindi and Kannada when using this fine-tuning approach.

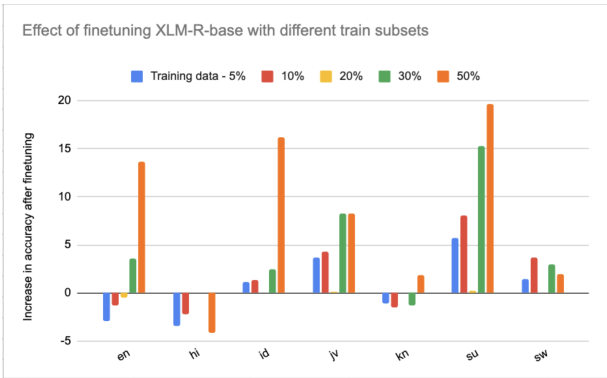

Figure 3: Hyperparameter study of how the performance of contrastive finetuning changes with different train/test splits

Swahili demonstrated a modest 3-4% improvement, further corroborating the potential benefits of contrastive fine-tuning for languages with some degree of relatedness or shared linguistic features. Interestingly, as the training data percentage increased to 50 %, we observed overfitting on the successful languages (Indonesian, Javanese, and Sundanese), which translated into further performance degradation for Hindi and Kannada. Based on these findings, we selected 30 % of the data for training, as this threshold did not drastically impair performance across languages while still yielding improvements for certain language groups.

5.2 Contrastive Finetuning Performance on XLM-R Models

Overall we observe high performance increase in sundanese (9-12%), Indonesian (5-6 %), Javanese (8-9%) and Swahili (3-4%). However we see minor (1-3%) performance degradation in Hindi and Kannada. We hypothesize that the substantial performance gains observed in these languages can be attributed to the geographical proximity of the regions where these languages are spoken, as well as the inherent linguistic similarities among them. Contrastive fine-tuning appears to effectively capture these patterns and facilitate cross-lingual knowledge transfer. Conversely, Hindi and Kannada, which employ distinct scripts and share fewer conceptual overlaps with the other languages in our dataset, exhibited a decrease or negligible improvement in performance. Future work could explore increasing the volume of data for Hindi and Kannada to mitigate the observed performance decline.

5.3 Chain of prompting on GPT models

After running CoT on GPT 3.5 turbo, we saw that the model performance increased for hi (20%) and su (15%), and decreased for en_dev (-7%), id (-6%), jv (-2%), kn (-6%), sw (-10%), and yo (-5%). The metrics are given in Table 4.

5.4 Cultural document augmentation on GPT models

Despite the introduction of a curated dataset, models processing Hindi (60.87%) and Kannada(54.6%) continue to struggle due to com-

plex grammar and cultural idiomatic expressions. Additionally, the nuanced script and phonetics further complicate comprehension. While the dataset aimed to bridge these gaps, its volume and variety were insufficient to capture the full linguistic subtleties. In contrast, language models perform well in Indonesian(83.27%) due to its simple grammar. However, this success hasn't extended to related languages like Javanese and Swahili, where interpreting metaphors requires deep cultural understanding.

6 Qualitative Analysis

6.1 Contrastive Finetuning for XLM-R models

6.1.1 Successful cases

Our qualitative analysis revealed several instances where contrastive fine-tuning significantly enhanced the model's performance across the dataset. One notable improvement was observed in the comprehension of simple metaphors and semantic nuances. For example, the initial XLM-R model failed to accurately interpret the metaphorical phrase "as encouraging as a dead bird" in English as implying a lack of support from a teacher. This failure likely stemmed from the model's learned association between teachers and positive sentiments, preventing it from appropriately prioritizing the negatively connoted "dead bird" phrase. Similarly, in the Javanese language, the improved model successfully associated the metaphor "Banyune kolam renang kaya banyu ler" (meaning "the swimming pool is like fresh water") with the concept of "bening" (clear). instead of incorrectly translating it to "buthek" (murky) like the original model.

Interestingly, we also observed improvements in the model's cultural understanding after contrastive fine-tuning. For instance, the fine-tuned model correctly interpreted the Indonesian phrase "bagaikan bias dihitung jari" (literally "counting on fingers") as implying "sedikit" (easy or few), whereas the original model lacked this cultural context.

These enhancements in the model's ability to comprehend unseen metaphorical expressions and cultural nuances suggest that contrastive fine-tuning effectively captures meaningful linguistic patterns, particularly when the training data contains similar associations. Furthermore, the improved performance on related languages indicates that conceptual knowledge transfers more effectively across languages through this fine-tuning approach.

6.1.2 Error cases

The qualitative analysis of failure cases highlights the finetuned XLM-R model's limitations, particularly in handling cultural references and concept shifts unique to certain languages. For example, the model struggles with Hindi, where the word for owl, "उल्लू," is associated with foolishness, a cultural nuance not present in its training data. Similarly, in Swahili, the phrase "fish out of water" translates to "not straight," illustrating the model's difficulty with cultural nuances and shifts, impacting its accuracy.

Moreover, the analysis reveals that the model still faces challenges in interpreting medium complexity examples in languages like Hindi and Kannada. In Hindi, the phrase "समुन्दर में मछली पकड़ने," which translates to "catching a fish," is associated with toughness instead of ease, indicating a lack of semantic understanding. Similarly, in Kannada, comparing a house to a hut, or "ಗುಡಿಸಲಿನಂತೆ" is incorrectly linked with a palace, highlighting semantic comprehension issues. Although contrastive finetuning has improved performance, further research is needed to address these limitations and enhance the model's multilingual capabilities.

6.2 Chain-of-Thought Prompting for GPT

After running CoT on GPT 3.5 turbo, we saw that the model performance increased for hi and su, and decreased for en_dev, id, jv, kn, sw, and yo. That was a surprising result. The hi and su languages are from different language families. Also, hi is from the Indian mainland whereas su is an island language. Su shares that with id and jv, but their performance decreased. Su has an English-like script but some other languages also have an English-like script but they didn't perform better. Hence, we still need to pinpoint exactly what sets hi and su apart from the other languages to cause the different performance. The reason could

potentially be something in the grammatical structure of the languages, such as hi, su, and kn have Subject-Object-Verb (SOV) order whereas the other languages have SVO. However, it is important to note that even though the % change in the accuracy can seem large in Table 4. We only tested CoT with one set of hyperparameters on GPT 3.5, where the temperature was the default value, the max_tokens were 150 and there were specific in context examples.

6.2.1 Failure of CoT

The model was still not able to perform well on examples which needed cultural references. As seen in Table 2 for the phrase "शंकर कार राजधानी की तरह चला रहा है।", which means "Shankar is driving his car like Rajdhani". Here Rajdhani is the reference to one of the major trains in India called Rajdhani Express. It travels very fast and stops at very few stations. The phrase "शंकर कार बहुत तेज चला रहा है।" (Shankar is driving his car very fast) is closer in meaning than "शंकर कार बहुत धीरे चला रहा है।" (Shankar is driving his car very slow). The explanation that the model gave was "Explanation: The metaphor comparing Shankar's car to the capital implies that Shankar's car is moving slowly and steadily, just like how the capital city moves in a predictable and methodical manner. Therefore, the option ""शंकर कार बहुत धीरे चला रहा है।"" is the closest meaning to the metaphor." The model is correct in that "Rajdhani" in Hindi also means capital city. But the word is not being used in that context here. Hence, it makes another leap to "predictable and methodical" and says that he is driving slowly. The second type of issue seen was the case where the model did not understand the vocabulary of the language correctly. E.g. check this explanation from the model: "Explanation: The metaphor comparing someone to a scale being straight implies that the person is honest, fair, and balanced in their actions, similar to how a scale needs to be straight to give accurate measurements. Therefore, the closest meaning is ""वो चालक है।"" (They are honest/fair).". In this case, the original metaphor is "वो स्केल की तरह सीधा है।" and the other option is "वो मासूम है।" which actually means that (he) is innocent/not crafty. Hence, the model was correctly able to translate the original metaphor and had the right idea about the answer, but was able to correctly match the option to the intended meaning. All these occurrences sound reasonable since the CoT prompt did not add any new information to the model. Hence, the model does not have access to cultural references and still has an existing low amount of pre-training for the vocabulary.

6.2.2 Success of CoT

An example seen in Table 3 where the model performed incorrectly previously but gave the correct answer after CoT was the metaphor "उसका गुस्सा बर्फ की सिल्ली सा है।" which means "His anger is like a slab of ice". The two options are "उसका गुस्सा जल्दी उतर जाता है।" (His anger cools down very easily) and "उसको गुस्सा बहुत ज्यादा आता है।" (He gets angry very easily). With CoT the model gives the correct explanation: "Explanation: The metaphor ""गुस्सा बर्फ की सिल्ली सा है"" is comparing the anger to ice that melts quickly. Option 0) ""उसका गुस्सा जल्दी उतर जाता है।"" is the most fitting explanation as it means the anger subsides quickly.". The model was potentially able to correct the initial incorrect prediction because asking it to explain caused it to consider the property of ice melting quickly and making the connection anger subsiding quickly. Other examples which were correct also hugely seem to have a pattern where the logical connection with the metaphor is established by asking it to consider the properties/qualities of entities in the metaphors. However, it seems to work in some cases but not in other cases. We have not been able to pinpoint the specific characteristics where CoT works in the above case.

6.3 Cultural Document Augmentation for GPT

The model continues to show challenges in interpreting metaphors that demand a deep understanding of cultural and contextual nuances, especially with Hindi and Kannada phrases. For instance, the Hindi metaphor "उसका गुस्सा बर्फ की सिल्ली सा है।" which means "His anger is like a big brick of ice" was interpreted by the model as indicating intense anger, overlooking the metaphorical nuance of being cold and unyielding, much like ice. Similarly, "उसका दामन काले कम्बल की तरह है।" which

Model	Input	True Meaning	Predicted Meaning	Reasoning
Finetuned XLM-R-base (hi)	उसने समुन्दर में मछली पकड़ने जैसा काम किया।	उसने बड़ा मुश्किल काम किया।	उसने बड़ा आसान काम किया।	Lack of semantic understanding
Finetuned XLM-R-base (sw)	Maneno yake yalikuwa kama samaki aliyetolewa kwa maji	Maneno yake hayakuwa ya-menyooka	Maneno yake yalikuwa yameny-ooka	Cultural concept shift
Finetuned XLM-R-large (kn)	ಆ ಕಟ್ಟಡವು ಗಡಿಯಾರದಂತೆ ಐಷಾರಾಮಿಯಾಗಿದೆ	ಆ ಕಟ್ಟಡವು ಐಷಾರಾಮಿಯಾಗಿಲ್ಲ	ಆ ಕಟ್ಟಡವು ಐಷಾರಾಮಿಯಾಗಿದೆ	Lack of semantic understanding
Finetuned XLM-R-large (hi)	कुछ लोग उल्लू की तरह बात करते हैं।	कुछ लोग लोग बड़ी बेवकूफी भरी बात करते हैं।	कुछ लोग अच्छी तरह बात करते हैं।	Culture-specific reference
Cultural Document Augmentation	उसका दामन काले कम्बल की तरह है।	वो दुष्ट है।	वो मासूम है।	Cultural Metaphor
Chain-of-thought (hi 1)	शंकर कार राजधानी की तरह चला रहा है।	शंकर कार बहुत तेज चला रहा है।	शंकर कार बहुत धीरे चला रहा है।	Cultural/geo-graphical refer-ence
Chain-of-thought (hi 2)	वो स्केल की तरह सीधा है।	वो मासूम है।	वो चालक है।	Poor vocabulary understanding

Table 3: Error Analysis of Proposed Methods

Table 4: Accuracy Comparison for GPT 3.5 on 100% of the data

Lang	Without CoT	With CoT	% change
en_dev	94.33	87.5	-7.24
hi	52.5	62.9	19.81
id	84.64	79.7	-5.84
jv	66.5	65.2	-1.95
kn	54.84	51.6	-5.91
su	47.33	54.5	15.15
sw	82.69	74.1	-10.39
yo	53.01	50.1	-5.49

means “His personality is like a black blanket” was misread as suggesting innocence and warmth, instead of implying hidden, darker aspects of a personality. In Kannada, the phrase "ಆ ಪ್ರದೇಶ ದೇವಸ್ಥಾನದಷ್ಟುಹಿಂಸೆಯಿದ ಕೂಡಿದೆ|" means “That region is filled with violence like a temple.” should have highlighted an ironic, contrasting level of violence, yet the model only recognized the direct notion of violence, missing the nuanced contrast with the peacefulness typically associated with a temple. Moreover, in Indonesian, the model’s

Table 5: Accuracy Comparison for Cultural prompting on GPT

Language	Initial	Cultural Doc	% Change
en_dev	94.33	93.41	-0.98
hi	52.5	60.87	-1.16
id	84.64	83.32	0.15
jv	66.5	64.72	-2.68
kn	54.84	54.6	-0.43
su	47.33	46.55	-0.16
sw	82.69	79.6	-3.74

interpretation of ”Aku pikir, kehidupan itu layaknya mengendarai mobil di jalan tol.” (I think life is like driving a car on a toll road) failed to capture the intended metaphorical depth, simplifying life’s journey to mere driving, thus misinterpreting the nuanced implications of navigating life’s complexities. These issues underline the model’s difficulties with complex language constructs and underscore

the necessity for enhancing its training on culturally rich, metaphorically dense content.

Some improvements in the model's performance in interpreting metaphors are largely due to the method of training with manually curated cultural documents, which enhances the model's contextual understanding and cultural sensitivity. This method equips the model with a deeper grasp of cultural nuances, as seen in the interpretation of the Hindi phrase "जीवन मीठा गुलकन्द है।" which means "Life is like a sweet gulkand", where life is compared to 'sweet gulkand', a beloved Indian preserve. The model now interprets this as Life is good, effectively capturing the essence of sweetness and contentment associated with gulkand. Similarly, In Sundanese, the phrase "Duit bisa ngarobah jalma kawas Qarun anu dikubur ku hartana sorangan." highlights the transformative power of money with a reference to Qarun, known for being buried by his own wealth. The model's response "Duit bisa ngarobah jalma jadi sarakah" (Money can turn a person into a beggar) captures the drastic impact of wealth, albeit with a slight deviation from the complete cultural narrative. These improvements demonstrate that the model can now appreciate and express the deeper emotional and cultural connotations embedded within metaphors, marking a significant advancement over earlier capabilities.

7 Future Work

As future work, we can explore contrastive finetuning with larger volume of data in hindi/kannada or to see if cultural specific knowledge can be learned while avoiding performance degradation with current test/train split. Instead of contrastive finetuning, we could also explore other losses like language modelling with a larger dataset with cultural information to understand how they help with performance with metaphors.

Some other things to explore in the future beyond this project are Few-shot CoT with examples and explanation of how they are figuratively different in the different languages. We can also provide the prompt in a language which is closer to the test language and test performance of both the encoder and decoder models. It would also be informative to codify the types of errors observed and make a comprehensive standard list and manually annotate each error to get a quantitative proof of how the issue types varying over languages and over methods.

8 Conclusion

In this report, we have conducted a comprehensive literature survey of the multi-cultural figurative QA and analysed multiple competitive baselines. We proposed three novel techniques to address the limitations. 1) Contrastive finetuning, which leverages contrastive learning to better capture figurative semantics; 2) Chain-of-Thought (CoT) prompting, which encourages models to reason step-by-step, aiding in contextual understanding; and 3) Document augmentation, which supplements the input with relevant cultural and linguistic information. Further, we performed a detailed error analysis of all the models, and proposed the strengths and weaknesses of our poposed methods with respect to different error categories. We found best performance increase with XLM-R models, for Id, Su, and Jv and our current state of in-context learning methods did not improve performance.

9 Contributions

- Shreya contributed to the model proposal, GPT-3 and XGLM model evaluation. In the model improvements, she worked on fine-tuning XLM-R base and XLM-R large.
- Anvesha contributed to the literature review, mBERT model evaluations. She curated the cultural documents for the languages and tested GPT 3.5 performance.
- Prachiti contributed to the XLM-base and XLM-large model evaluations in replications. In the improvement efforts, she worked on GPT-3.5 Chain-of-thought.
- All of us contributed to the report.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages.

Workshop BigScience, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn

de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Yunxiang Zhang and Xiaojun Wan. 2022. Birdqa: A bilingual dataset for question answering on tricky riddles.

Language	**mBERT-base**	**XLM-R-base**	**XLM-R-large**	**GPT-3.5**	**XGLM-large**
en_dev	65.6	73.4	80.4	**94.33**	81.05
hi	51.8	59.3	**66.5**	52.5	57.7
id	52	67.7	74.89	**84.64**	69.33
jv	51.33	52.05	59.66	**66.5**	61.33
kn	53.5	51.25	**59.17**	54.84	56.4
su	52.5	53.33	**59.66**	47.33	55.32
sw	52.36	52.8	59.09	**82.69**	56.17
yo	51.18	-	-	**53.01**	-

Table 6: Zero shot Model performance on different languages

Model	Input	True Meaning	Predicted Meaning	Reasoning
mBERT(hi)	राम के अभिषेक की बात सुनकर कौशल्या का अंग-अंग सूखने लगा।	राम के अभिषेक की बात सुनकर कौशल्या बहुत दुखी हुई।	राम के अभिषेक की बात सुनकर कौशल्या बहुत प्रसन्न हुई।	Uncommon concept globally
mBERT(hi)	उसके ऊपर अमृत सवार है।	वह बहुत प्यार बाँट रहा है।	वह बहुत प्यार बाँट रहा है।	Correct
mBERT(kn)	ಆ ಹಣ್ಣು ಹಲಸಿನಂತೆ ಬಹಳ ಅಪರೂಪ	ಆ ಹಣ್ಣು ಬಹಳ ಅಪರೂಪ	ಆ ಹಣ್ಣು ಬಹಳ ಅಪರೂಪವಲ್ಲ	Poor Understanding
mBERT(jv)	Pendukunge Pak Rangga dadi bupati kayata jamaah sholat subuh ing masjid	Pendukunge Pak Rangga dadi bupati sitik	Pendukunge Pak Rangga dadi bupati Akeh	Cultural Reference
XLM-R(id)	Tetangga baru gayanya setinggi langit.	Tetangga baru banyak gaya.	Tetangga baru mati gaya, norak, kampungan.	Poor Understanding
XLM-R(kn)	ಆ ಅವನು ದುಃಖಿತನಂತೆ ಆಸೆಗಳನ್ನು ತ್ಯಜಿಸಿದ್ದನು	ಆ ಅವನು ಆಸೆಗಳನ್ನು ತ್ಯಜಿಸಿಲ್ಲ	ಅವನು ಆಸೆಗಳನ್ನು ತ್ಯಜಿಸಿದ್ದನು	Cultural knowledge
XLM-R(su)	ceuk ema jadi jalmi teh kedah sing someah hade ka semah	ceuk ema jadi jalmi teh kedah ramah	ceuk ema jadi jalmi teh kedah judes	Language comprehension issue
GPT 3.5 (kn)	ಆ ಸಮುದ್ರ ಕಬ್ಬಿಣದಂತೆ ತಣ್ಣಗಿದೆ	ಆ ಸಮುದ್ರ ತಣ್ಣಗಿದೆ	ಆ ಸಮುದ್ರ ತಣ್ಣಗಿಲ್ಲ	Poor understanding, literal case
GPT 3. 5(hi)	उसके साथ बहस करना हथगोले से खलने जैसा है।	उसके साथ बहस करना बहुत मुश्किल है।	उसके साथ बहस करना बहुत आसान है।	Figurative knowledge limitation
GPT-3.5 (hi)	बैंक में कुँए जितना पैसा भी नहीं है।	बैंक में बहुत कम पैसा है।	बैंक में बिलकुल पैसा नहीं है।	Minute cultural meaning change.
XGLM (sw)	Upendo ni kama kumwaga maji baharini	Upendo hauna umuhimu	Upendo ni muhimu	Incorrect interpretation
XGLM (id)	Youtuber itu makan nasi padang seolah perutnya sudah penuh.	Youtuber itu makan nasi padang dengan porsi sedikit.	Youtuber itu makan nasi padang dengan porsi sangat banyak.	Language comprehension limitation
XLM-R(id)	Tulisan tangannya bagaikan milik dokter.	Tulisan tangannya berantakan.	Tulisan tangannya rapi.	Cultural knowledge (doctor's handwriting)

Table 7: Error Analysis