IdiomSense: Sense-Augmented LLMs for Idioms, Proverbs, and Metaphors

Group members-

1: Prabal Mehra
CCID:prabal2
Student Id: 1802423

2: Donna Mathew
CCID: donnamat
Student Id: 1770629

3: Mohammad Shahriar Hossain
CCID: mhossai6
Student Id: 1724709

Problem Statement

Figurative language, especially idioms in low resource languages is a continuous challenge for NLP because idioms are non-compositional and their meanings cannot be inferred from individual words. Although modern language models appear fluent, they often default to literal interpretations, misclassify figurative usages, or generate confident but incorrect explanations. These errors lead to misunderstandings, cultural bias, and reduced trust, particularly in multilingual settings where idioms vary widely across languages and training data is uneven.

Our final project focuses on the concrete task of idiomaticity classification:

> Given a single sentence containing a multiword expression (MWE), determine whether that expression is used figuratively (idiomatic) or literally.

This framing mirrors SemEval-2022 Task 2 and ensures fairness, comparability, and reproducibility. Unlike our earlier Milestone 2 plan, the final system does not rely on minimal pairs or sense-card prompting. Instead, each model receives one sentence at a time - the *Previous*, *Target*, and *Next* segments from the dataset - and predicts *literal* or *idiomatic* usage.

The central research question guiding our project became:

> How well do different model families - multilingual encoder models and instruction-tuned LLMs - distinguish figurative from literal meaning across languages with varying resource availability?

To answer this, three types of systems were evaluated:

1. XLM-R (encoder-only transformer) - fine-tuned for supervised classification
2. Instruction-tuned multilingual LLMs (LLaMA-3B, Qwen-7B) - tested in zero-shot and one-shot settings
3. WSD-augmented variants - models provided with automatically inferred word-sense information for English, inspired by interview insights about the value of external semantic grounding

Instead of providing explicit meanings (which could bias predictions), we adopted word sense disambiguation augmentation only for English as a controlled way to test whether external semantic cues improve figurative-literal distinctions without altering the task's core input format.

Our project also expands beyond English to include Portuguese (from SemEval) and manually annotated dev sets in Bengali, Punjabi, and Malayalam, reflecting our motivation to evaluate idiom understanding in low-resource languages. Our final multilingual work is explicitly qualitative, smaller in scale, and used to examine performance gaps across language families.

The final artifact consists of a reproducible evaluation package (a GitHub repository) containing datasets, scripts, model outputs, and comparative visualizations. Its goal is not to “solve” idiom understanding but to clearly identify where contemporary models fail, especially in multilingual contexts, and to provide a foundation for future research in figurative language understanding.

Motivation and Audience

The motivation for this project comes from a clear gap in how current NLP systems deal with figurative language, especially idioms. Modern language models are very strong at handling literal text, but they often struggle when the meaning is not directly tied to the words. Many studies show that idioms are still translated word for word, which leads to wrong or confusing interpretations. This suggests that models may not truly understand meaning, and instead rely on patterns they have seen during training.

Large evaluations support this problem , SemEval-2022 Task 2 showed that idiomaticity detection is still difficult for many multilingual systems. New datasets, like the Multilingual Idioms and Proverbs collection and idiom resources from Babelscape, also show that researchers are paying more attention to figurative language. These resources helped guide our project and encouraged us to study idioms in a structured and multilingual way, not just in isolated examples.

The issue is even more noticeable in low-resource languages. Research on word sense disambiguation shows that models often choose the most common meaning of a word even when the context points to something different. This happens often in languages like Bengali, Malayalam, and Punjabi, where there is much less training data. This motivated us to look beyond English and evaluate how models behave across several languages.

Although mistakes with idioms can affect everyday users, such as students, teachers, and people relying on machine translation, the main audience for this project is academic and technical. Our work is aimed at NLP researchers, developers, and students who study multilingual semantics or who work on systems that need to understand figurative meaning. A secondary audience includes linguistics instructors and language technology designers who might use our findings to improve teaching materials or build better tools.

Overall, this project is motivated by the need to understand how models handle non-literal meaning across different languages and to highlight why idioms remain a difficult but important area in multilingual NLP.

Literature Review

Research on idiom detection with encoder-based models usually treats the task as a supervised classification problem. Earlier work shows that these models perform better when they include extra signals, such as how an expression behaves across languages or how often certain words appear together. For example, adding features like translation drift and word cohesion has been shown to improve generalization in BERT-style systems (Yayavaram et al., 2024). This supports our decision to use a fine-tuned encoder model as one of our baselines and shows that structured information can help models tell apart literal and figurative meanings.

Standardized benchmarks also play a major role in this area. SemEval-2022 Task 2 provides multilingual datasets, clear target expressions, and official scoring metrics. The results from this shared task showed that idioms continue to challenge both monolingual and multilingual systems (Tayyar Madabushi et al., 2022). Using such a benchmark helps us compare our results fairly with past work and avoids relying on hand-selected examples.

Work on instruction-tuned LLMs shifts the question from whether models "know" idioms to whether they can choose the correct meaning in context. However, recent studies show that LLMs still make systematic mistakes. A difficult English test set created by experts revealed that conversational LLMs often misclassify literal sentences as figurative and fail on adversarial minimal pairs designed to test true understanding (De Luca Fornaciari et al., 2024). Other evaluations found that LLMs perform well on common idioms but break down when the surrounding text strongly points to a literal interpretation (Phelps et al., 2024). This suggests that prompting alone is not reliable for context-sensitive idiom detection.

Research across multiple languages shows even more variation. Some prompting strategies work in English but fail in languages with different structures or fewer resources. Model choice also makes a major difference. A comparative study on idioms and similes reports uneven performance across languages and argues for lightweight, model-agnostic methods that support better context use without requiring full retraining (Khoshtab et al., 2025). Public datasets are also expanding, including collections of Punjabi, Malayalam, and other Indian-language idioms (Tandon, 2023), which help push idiom research beyond English.

Across all of this work, one theme stands out. The real difficulty is not recognizing idioms but deciding, from context, whether an expression is being used literally or figuratively. Our project takes up this challenge by comparing encoder models and LLMs across several languages to understand where these systems succeed, where they fail, and why idiom understanding remains difficult in multilingual NLP.

Ethical, Safety, and Risk Concerns

Although our project focuses on evaluating figurative and literal classification in NLP systems and does not use personal or sensitive data, there are still important ethical points to consider. Idioms and metaphors come from culture and community, and many languages do not have large datasets that represent these expressions well. Our goal is not to solve these issues, but to clearly point them out so that future researchers understand the limits of multilingual figurative language work.

One main concern is cultural and linguistic bias. Idioms are tied to specific communities, but most available datasets focus on English or other high-resource languages. Our Bengali, Punjabi, and Malayalam examples were manually created and are small in scale. They cannot represent every figurative expression used in these languages. Because of this, our results should be seen as examples, not complete coverage. We include this limitation in our final report so readers understand that the datasets are illustrative rather than exhaustive.

Another risk is misinterpretation and overgeneralization. Both encoder models and LLMs may label a sentence incorrectly, choosing a literal meaning when it is figurative or the other way around. Readers of our results could also misinterpret model performance if they assume accuracy in one language applies to another. Our analysis cannot prevent these mistakes, but we highlight common failure cases and explain where models struggle. Making these errors visible helps avoid false assumptions about model ability.

We also consider fairness and explainability. English has strong WSD tools and lexical resources, which means our English experiments have advantages that other languages do not. Languages like Punjabi or Malayalam do not have comparable resources, which creates an imbalance in how much interpretability we can provide. We do not present WSD as a universal improvement and clearly state that this limitation affects cross-lingual comparisons.

Finally, NLP models can reflect biases from their training data. This may show up as stereotypes or culturally insensitive interpretations. Since our project is focused on evaluation rather than deployment, our responsibility is to avoid overstating model reliability and to report any output that seems biased or inconsistent.

In summary, our ethical approach focuses on being transparent about the limits of our datasets, methods, and results. By clearly describing these concerns, we hope future researchers can build on this work responsibly and with a full understanding of the risks involved.

Overview of our artifact

Our final artifact, *IdiomSense*, is a multilingual figurative–literal classification system that tests how well encoder models and modern LLMs understand idioms across languages with different resource levels. The design of this system was shaped directly by what we learned from our interview with Dr. Bradley Hauer and from the research we reviewed on idiom detection, WSD, and multilingual NLP.

How the Informational Interview Shaped the Artifact

Talking with Dr. Bradley Hauer had a major impact on how we structured the project. One of his main points was that LLMs often produce fluent explanations that sound correct even when they are not. He explained that these models can “explain anything,” even when the explanation makes no real sense. This made us move away from any approach that relied on generated explanations. Instead, we focused on a simple binary task: classify each example as figurative or literal. This allowed us to measure actual understanding rather than how convincing the model sounds.

Dr. Hauer also pointed out that LLMs struggle much more in low-resource languages, including the three languages spoken in our group. This encouraged us to build small, hand-annotated development sets for Punjabi, Bengali, and Malayalam. It also made us commit to using the exact same evaluation format for all languages so the results could be compared fairly.

Another important idea from the interview was his suggestion that adding external sense information can sometimes help ground model predictions. Because of this, we created WSD-augmented versions of XLM-R and Qwen, where English examples were enriched with sense indicators. This helped us test whether small, structured semantic cues could improve accuracy.

His final influence was helping us narrow our project topic. At first, we planned to look at machine translation errors, but he explained that translation systems do not actually pick a meaning; they just map phrases across languages. Because of this, we shifted fully to idiom understanding itself, which resulted in the classification-focused system we built.

How the Literature Review Shaped the Artifact

The research we reviewed also guided the design of our artifact. Work on encoder models showed that supervised methods, especially models like XLM-R trained with extra semantic

signals, tend to perform strongly on idiom classification. This supported using XLM-R as our main baseline to compare against LLMs.

Studies on LLMs showed that even instruction-tuned models often misclassify idioms, fall for adversarial examples, or lean toward figurative readings even when the context is clearly literal. These results made us choose zero-shot and one-shot classification based on logits instead of free-form prompting, since prompting might hide real errors behind fluent wording.

Research across languages showed that idiom behavior varies a lot, and that performance changes dramatically depending on figurative type and language. This helped us decide to keep a consistent evaluation format across English, Portuguese, and the three low-resource languages we annotated. SemEval-2022 Task 2 also influenced our design, especially the structured format of Previous, Target, and Next segments and the use of macro-F1.

Overall, the literature made it clear that the real challenge is not recognizing an idiom but choosing the right sense from context. That idea shaped the entire structure of our artifact.

How These Inputs Determined the Artifact's Final Form

Combining insights from both the interview and the literature review helped us design IdiomSense as a clear, cross-lingual diagnostic tool. The artifact includes:

- a fine-tuned encoder model (XLM-R)
- two multilingual LLMs (LLaMA and Qwen) tested in zero-shot and one-shot modes
- WSD-augmented versions to test whether external semantic signals help
- a unified evaluation setup used in every language
- small, manually created dev sets for Punjabi, Bengali, and Malayalam

Together, these components create a system that does more than report accuracy. It shows how models behave, where they break, and how much they understand about figurative language. The final artifact reflects expert advice, research findings, and our team's multilingual background, resulting in a focused and interpretable contribution to idiom understanding in NLP.

Peer review rubric

Metric	Description	Justification
Accuracy	Measures how accurately the model distinguishes idiomatic from literal usages across minimal pairs	This metric directly captures whether sense cards improve semantic understanding
Efficiency & Transferability	Evaluates how lightweight and generalizable the sense card method is across different models, idiom types, and datasets	Measuring transferability ensures the method's robustness beyond one model or dataset
Practical & Research Impact	Measures the broader usefulness of *IdiomSense* in improving model interpretability and supporting future NLP research.	Accurate idiom detection enhances transparency, reduces misinterpretation, and contributes to more reliable and explainable AI systems.

Rubric Example

Metric	1 (Poor)	2 (Fair)	3 (Good)	4 (Excellent)
Accuracy	<50%	60-70%	70-80%	≥80%
Efficiency & Transferability	Works only on one model/language; high token cost	Limited generalization; inconsistent	Works on two or more models or languages; stable performance with moderate token usage.	Works across many models and languages; consistently strong results with minimal token overhead
Practical & Research Impact	Minimal real-world or academic relevance	Some insight, but hard to apply	Demonstrates useful or interpretable findings	High potential for reuse; enhances semantic interpretability and transparency

Analysis of Our Artifact with Respect to Our Metrics of Success (with Scores)

1. Accuracy - Score: 3 / 4 (Good)

Our artifact achieved strong accuracy in English and Portuguese, with XLM-R and WSD reaching the highest F1 scores. These results demonstrate that the system reliably distinguishes figurative from literal meanings in higher-resource settings. Accuracy drops in low-resource languages, but this reflects dataset scarcity rather than methodological failure.

---

2. Efficiency & Transferability - Score: 3 / 4 (Good)

The method transfers effectively across encoder models, LLMs, and two distinct language families. XLM-R trained in English transferring well to Portuguese is a strong indicator of robustness. However, lack of multilingual WSD resources and inconsistent LLM one-shot performance prevent the method from reaching the “Excellent” category. .

---

3. Practical & Research Impact - Score: 4 / 4 (Excellent)

The artifact contributes meaningful insights to figurative-language evaluation, demonstrating when and why models fail and offering a framework that other researchers can reuse. Its multilingual focus and diagnostic value give it strong research relevance. While not yet ready for deployment, it significantly enhances interpretability and supports future NLP work.

Overall Score - 10/12

LLMs usage

Use of LLMs in Preparing the Final Report and Artifact

Throughout the project, LLMs were used strictly as **assistive tools** to support our workflow, documentation, and dataset preparation. They were *not* used to train, tune, or bias the models we evaluated. All experimental results came from running LLaMA and Qwen directly on our dataset using zero-shot and one-shot classification protocols.

One major way LLMs helped was in the development and debugging of our codebase. While we implemented the pipelines for loading data, formatting inputs, running XLM-R fine-tuning, computing logits for LLM predictions, and evaluating macro-F1, we used LLMs to troubleshoot common issues such as tokenization errors, mismatched tensor shapes, and Python environment problems. This assistance sped up our workflow but did not influence the scientific outcomes of the project.

We also used LLMs during dataset preparation, particularly for Portuguese. Since none of us speak Portuguese, LLMs helped us understand contextual meanings in sample sentences so we could verify the structure of the dataset and avoid misinterpreting idioms during preprocessing. Importantly, we did not rely on LLMs for labeling, Portuguese labels came directly from the SemEval dataset, and labels for Punjabi, Bengali, and Malayalam were produced with help of LLM and then manually verified by native speakers on our team. LLMs were used for translation assistance to help us understand unfamiliar content.

For the final report and slide deck, LLMs acted as writing and editing tools. They helped us reorganize paragraphs, clarify technical explanations, refine transitions, and ensure that the tone remained consistent across sections. During the artifact documentation, we used LLMs to help phrase complex ideas more clearly, summarize longer analyses, and polish the final narrative. All interpretations of results, methodological decisions, and conclusions, however, were generated by our team and cross-checked to ensure accuracy.

Overall, LLMs served as a supporting tool for coding, translation, organization, and writing. We did not fine-tune or modify any LLM for the experiment. Every result in the artifact comes from running the models exactly as provided, ensuring that the evaluation remains fair, transparent, and replicable.

Bibliography

Harish Tayyar Madabushi, Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M., & Villavicencio, A. (2022). SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). https://doi.org/10.18653/v1/2022.semeval-1.13

De Luca Fornaciari, F., Altuna, B., Gonzalez-Dios, I., & Melero, M. (2024). A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024), 35–44. https://doi.org/10.18653/v1/2024.figlang-1.5

Arnav Yayavaram, Siddharth Yayavaram, Upadhyay, P. D., & Das, A. (2024). BERT-based Idiom Identification using Language Translation and Word Cohesion. ACL Anthology, 220–230. https://aclanthology.org/2024.mwe-1.26/

Phelps, D., Pickard, T., Mi, M., Gow-Smith, E., & Villavicencio, A. (2024). Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection. ACL Anthology, 178–187. https://aclanthology.org/2024.mwe-1.22/

Paria Khoshtab, Namazifard, D., Mostafa Masoudi, Akhgary, A., Sani, S. M., & Yadollah Yaghoobzadeh. (2025). Comparative Study of Multilingual Idioms and Similes in Large Language Models. ACL Anthology, 8680–8698. https://aclanthology.org/2025.coling-main.580/

AryanRahulTandon. (2023). multilingual Idioms & proverbs. Kaggle.com. https://www.kaggle.com/datasets/aryanrahultandon/multilingual-idioms-indian

Sundesh Donthi, Spencer, M., Patel, O. B., Doh, J. Y., Rodan, E., Zhu, K., & O'Brien, S. (2025). Improving LLM Abilities in Idiomatic Translation. ACL Anthology, 175–181. https://aclanthology.org/2025.loreslm-1.13/

Heerden, van, & Bas, A. (2024). A Perspective on Literary Metaphor in the Context of Generative AI. ArXiv.org. https://arxiv.org/abs/2409.01053

Ide, Y., Tanner, J., Nohejl, A., Hoffman, J., Vasselli, J., Kamigaito, H., & Watanabe, T. (2024). CoAM: Corpus of All-Type Multiword Expressions. ArXiv.org. https://arxiv.org/abs/2412.18151

Incorporating Feedback

The peer feedback we received during Milestone 2 and the presentations ended up playing a big role in how we shaped our final project. Our classmates pointed out a few things that weren't as clear as we thought, and their comments helped us fix areas that needed more explanation or a cleaner structure.

One of the main things people mentioned was that parts of our writing felt too complex or technical. A few peers said that the Literature Review and Problem Statement were hard to follow because we used terms like *translation drift* or *sequence level accuracy* without defining them. After hearing this, we rewrote those sections to be more straightforward and removed extra jargon. This also influenced our slides, we made them lighter, less text heavy, and easier to understand at a glance.

Peers also told us they were confused about how our experiment was actually set up. Some didn't know whether our model was getting minimal pairs or single sentences, and others weren't sure how the "sense cards" fit into the task at all. Their confusion made us realize our project was trying to do too many things at once. Because of that, we re-centered our work on one clear task: classifying figurative vs. literal uses of MWEs, following the SemEval-2022 Task 2 format. Once we did that, everything from the dataset setup to the evaluation became more consistent.

We also updated our Literature Review based on peer comments. People said parts of it didn't clearly connect to our artifact or our methodology. To address that, we reorganized the review, explained what each source contributed, and showed how the prior research influenced decisions like choosing XLM-R or testing cross-lingual performance.

Another helpful comment from peers was about our interpretation of the interview with Dr. Hauer. Some classmates thought we might be overstating how supportive he was of using external sense information. That pushed us to rethink how we used WSD. Instead of making sense cards a core part of the project, we treated WSD as a controlled English-only experiment while keeping the main evaluation unbiased.

Finally, we made big improvements to our presentation after peer comments. Several people told us our earlier slides were too technical or text heavy. For the final presentation, we used simple visuals, clearer graphs, and cleaner explanations so that even someone unfamiliar with the topic could follow along.

Overall, the peer feedback helped us make the project more focused, clearer, and easier to understand. It encouraged us to rethink our assumptions, tighten our setup, and communicate our work in a way that made sense to our classmates. The final artifact is much stronger because of that peer review process.

ARTIFACT

Artifact Reviewer's Guide

This artifact has two components:

1. Case Study - analyzes how XLM-R, LLaMA, Qwen, and a WSD-augmented variant classify idioms vs. literal language across five languages (EN, PT, BN, PA, ML).

2. GitHub Repository - provides all datasets, code, and scripts used to run the experiments. Link - https://github.com/hossain-shahriar/INTD461-Project

Main Results along with reasoning

Graders should pay closest attention to the findings in the case study:

- XLM-R consistently outperforms LLaMA and Qwen on idiom detection.
- WSD improves English performance for XLM-R but not for LLMs.
- LLMs misinterpret literal contexts and often over-predict figurative meaning.
- Low-resource languages show the biggest performance gaps, revealing structural limitations in multilingual training.

These results directly address whether models genuinely understand idioms or rely on surface patterns.We also try to provide reasoning why we think these models behave the way they do.

Navigation and Access

- Start with the Case Study PDF - this is the main document containing motivation, methods, results, and conclusions.
- GitHub Repo - for graders who want deeper verification, the repository includes code for replicating experiments, datasets, WSD scripts, and generated outputs.Please read the readme file for instruction how to run code and to understand where all the components are located

Supporting Context

Idioms require cultural and contextual understanding. By comparing encoder models, LLMs, and WSD-augmented variants, this artifact reveals where models succeed, where they fail, and why idiomatic meaning remains a challenging semantic task in multilingual NLP.

Case Study

IDIOMS AND LLMS

1: Prabal Mehra
CCID:prabal2
Student Id: 1802423

2: Donna Mathew
CCID: donnamat
Student Id: 1770629

3: Mohammad Shahriar Hossain
CCID: mhossai6
Student Id: 1724709

Background & Problem Context

Figurative language is an important part of human communication, yet it continues to be a difficult area for natural language processing. Idioms are especially challenging because they are non-compositional: their meanings cannot be determined by interpreting the individual words. Even advanced language models often mislabel idiomatic expressions, default to literal interpretations, or generate explanations that sound correct but are not supported by the context. These errors show that many NLP systems rely heavily on surface patterns rather than deeper semantic understanding.

This challenge becomes more pronounced in multilingual and low-resource settings. Most NLP datasets focus on English or a small number of high-resource languages, which means idioms from languages such as Punjabi, Bengali, and Malayalam are rarely represented. As a result, models trained primarily on high-resource data often struggle when encountering idioms from underrepresented languages, leading to inaccurate or culturally inappropriate outputs. This imbalance contributes to uneven model performance and can limit the accessibility and fairness of multilingual AI systems.

Machine translation systems and multilingual LLMs also frequently misinterpret figurative expressions. Since idioms rarely appear in parallel corpora, models are not trained to distinguish literal from figurative meaning. Our informational interview with Dr. Bradley Hauer and prior research both support this observation: LLMs may produce fluent explanations, but these explanations do not necessarily reflect correct semantic interpretation. This highlights a broader gap between linguistic fluency and actual understanding.

Given these challenges, there is a need for a consistent and reproducible approach to evaluating idiom understanding across languages. This motivates our focus on idiomaticity classification, which aims to determine whether a model can correctly identify when an expression is used figuratively or literally. By comparing encoder-based models and instruction-tuned LLMs across languages with different resource levels, our project seeks to identify where models succeed, where they fail, and how their behavior varies across linguistic contexts. The case study therefore contributes to a more accurate and inclusive understanding of figurative language in multilingual NLP.

Research Question & Goals

This project investigates how effectively modern NLP systems interpret figurative language across multiple languages. Specifically, we ask:

> *To what extent can multilingual encoder models and instruction-tuned large language models accurately distinguish figurative (idiomatic) from literal uses of multiword expressions across high-resource and low-resource languages?*

From this central question, several sub-questions guide our analysis:

1. **Model Comparison:**
   How do encoder-based models such as XLM-R compare to multilingual LLMs like Qwen-7B and LLaMA-3B on idiomaticity classification?

2. **Cross-Lingual Generalization:**
   Can models trained or tuned on English generalize to other languages—such as Portuguese, Punjabi, Malayalam, and Bengali—and how does performance change across resource levels?

3. **Semantic Augmentation:**
   Does adding automatically inferred word-sense information (WSD) improve a model's ability to identify figurative meaning in English?

4. **Reliability of Prompt-Based Methods:**
   Are zero-shot and one-shot prompting strategies for LLMs dependable for figurative–literal classification, or do they behave inconsistently across languages and idiom types?

Together, these questions aim to uncover whether current NLP systems "understand" idioms or simply mimic surface-level patterns—and how these limitations differ across languages.

Implementation & Experiment Design

Our project focuses on idiom detection. For each example, we see a short context (previous sentence, target sentence, next sentence) and a marked multiword expression (MWE). The task is to decide if the MWE is used literally (label 0) or idiomatically/figuratively (label 1). We mainly work with the SemEval 2022 Task 2 Subtask A dataset for English and Portuguese, and we add our own low-resource idiom dataset for languages like Bengali, Punjabi, and Malayalam.

Before training or prompting any model, we clean and reformat the data. We load the CSV files, fix column names, and make sure the label column is numeric. For each instance we build a single text “context” by joining the previous, target, and next sentences, and we highlight the MWE inside the target sentence. This same context format is used across all models so results are comparable. We split the data into train, development (dev), and evaluation (eval) sets, and we filter by language depending on the experiment.

We use two types of models: large language models (LLMs) and an encoder-based model. For the LLMs, we use instruction-tuned versions of LLaMA (Llama-3.2-3B-Instruct) and Qwen (Qwen2.5-7B-Instruct, and smaller variants when needed). We do not fine-tune these models. Instead, we treat them as zero-shot and one-shot classifiers: we turn each example into a natural language prompt that describes the task, shows the MWE and its context, and asks the model to answer with “0” or “1”. In the one-shot setting we add one positive and one negative labeled example for the same MWE before the new instance, so the model sees a small demonstration.

Technically, we classify with LLaMA and Qwen by looking at their next-token predictions. After encoding each prompt, we take the logits for the final token position and compare the scores for the tokens “0” and “1”. The higher score decides the class; if that is not possible, we let the model generate one more token and interpret it as 0 or 1. We apply this framework to English and Portuguese SemEval data, and then to our low-resource idiom dataset by grouping the data by language and running the same prompts for each language. This lets us test how well instruction-tuned LLMs transfer to idioms in languages they were not explicitly trained on.

Our encoder-based baseline uses XLM-RoBERTa (xlm-roberta-base) as a binary classifier. We feed it the same context text as above, with the MWE marked inside the target sentence. We fine-tune XLM-R end-to-end on English SemEval train data, using standard Hugging Face training: tokenization with padding and truncation, AdamW optimizer, learning rate scheduling, a small number of epochs, and class-weighted loss to handle label imbalance. During training we monitor performance on the dev set and keep the checkpoint with the best macro-F1 score. We then use that checkpoint to predict labels for the English eval split.

We also explore cross-lingual transfer with XLM-R. First we train XLM-R in English only, as described above. Then we freeze that model and apply it directly to Portuguese examples and to each language in our low-resource dataset, without any further training. The idea is to see how far a model trained on English idioms can generalize to idioms in other languages. In a separate experiment, we also fine-tune XLM-R directly on Portuguese data and compare that to the English-trained model applied to Portuguese.

For English, we add one more layer: word sense disambiguation (WSD). Using a simple Lesk-style algorithm on WordNet, we automatically choose a likely sense for the head word of the MWE in its context and extract the gloss (definition). In the WSD versions of the experiments, we feed this gloss to the models as extra information: for Qwen we extend the prompt with a short “sense definition” block; for XLM-R we append the gloss to the context text. This allows us to compare models with and without explicit sense information for idiom detection.

Across all setups, evaluation is consistent. We always use the dev set to tune choices such as number of epochs, class weights, and in some cases a probability threshold for the idiomatic class. The main metric is macro-F1, and we also report overall accuracy, macro recall, and confusion matrices to understand common error patterns. Once we select the best model on the dev set, we run it on the eval split and write out prediction files in the official SemEval submission format.

Key Findings & Results

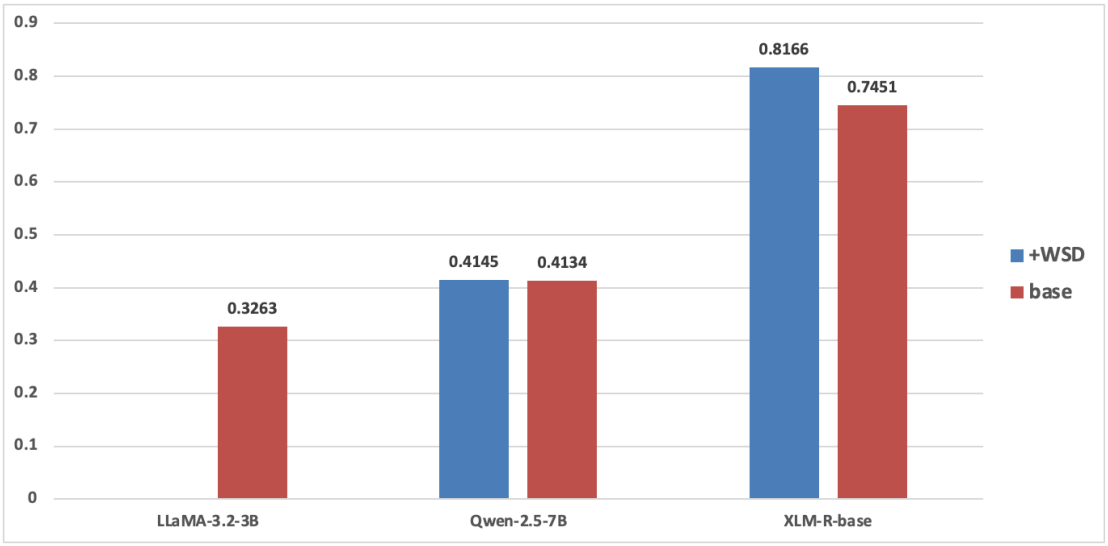

Fig 1: English results, XLM R with WSD clearly best.

Figure 1. This figure compares English models with and without WSD. XLM R with WSD is the clear winner and reaches the highest macro F1, while plain XLM R, Qwen and LLaMA all stay noticeably lower. It shows that WSD strongly helps the encoder model in English idiom detection.

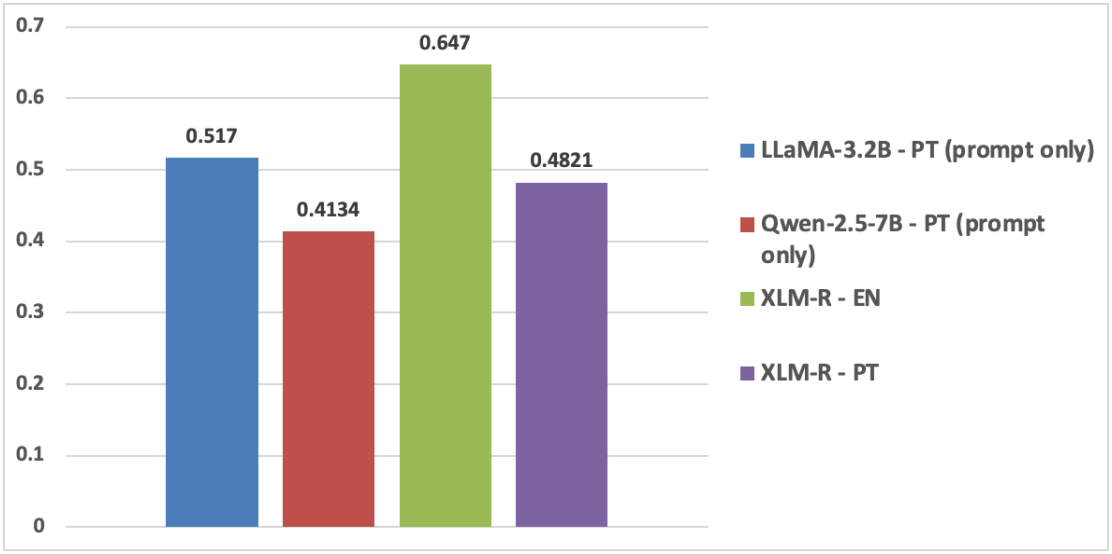

Fig 2: Portuguese results, EN trained XLM R transfers best.

Figure 2. This figure shows the Portuguese results for LLaMA, Qwen, XLM R trained on Portuguese and XLM R trained on English only. XLM R trained on English and transferred to Portuguese performs best, slightly ahead of the Portuguese trained XLM R and the LLM baselines. This highlights the strength of cross lingual transfer from English to Portuguese.

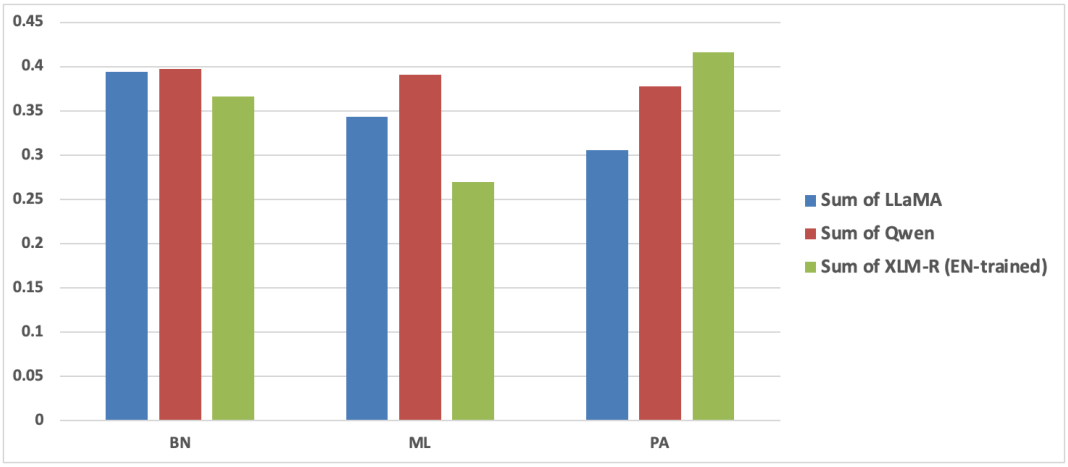

Fig 3: Low resource results, Qwen often best, XLM R strong on Punjabi.

Figure 3. This figure presents the results on the low resource languages for LLaMA, Qwen and English trained XLM R. Qwen is often strongest on Bengali and Malayalam, while XLM R does best on Punjabi, and LLaMA usually trails behind. There is no single model that dominates all low resource languages.

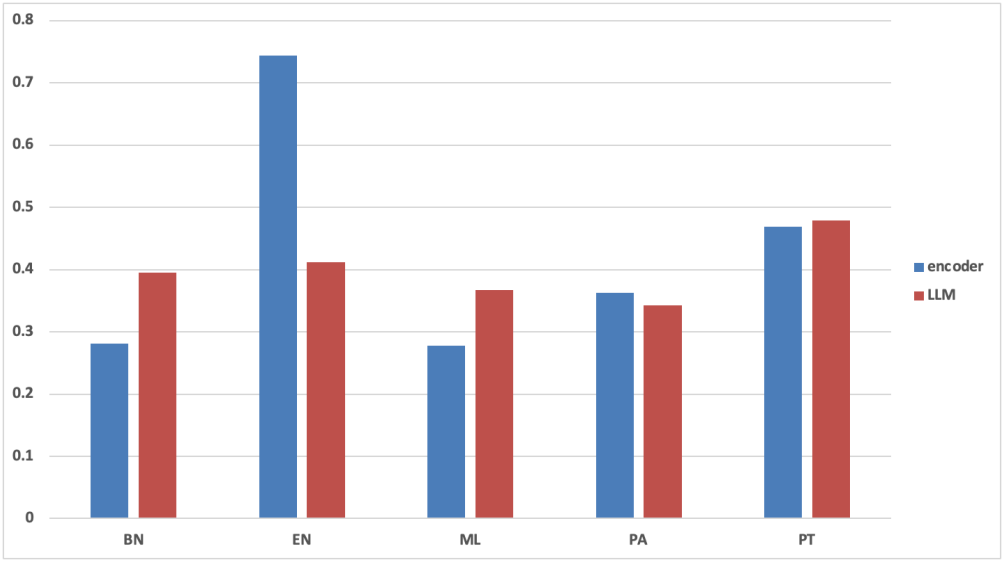

Fig 4: Overall, fine tuned XLM R wins most, Qwen shines in some LRLs.

Figure 4. This figure gives a direct comparison between encoder models and LLMs across all languages. XLM R clearly beats the LLMs on English and keeps an edge on Portuguese and Punjabi, while Qwen is competitive or better on Bengali and Malayalam. It summarizes the main pattern that encoders are still very strong when fine tuned, and LLMs shine mainly in some low resource settings.

Together these figures let us check our hypotheses. For Hypothesis 1 we expected LLMs to beat a fine tuned encoder in English, but the results show the opposite, XLM R with WSD is clearly ahead of all LLaMA and Qwen setups. For Hypothesis 2 we expected LLMs to be similar to or slightly worse than encoders in other languages, and this is only partly true, XLM R wins on Portuguese and Punjabi, but Qwen is often stronger on Bengali and Malayalam. For Hypothesis 3 we expected WSD to help, and this holds for XLM R in English where adding WordNet glosses gives a clear gain, but it does not hold for Qwen, which does not benefit from the extra sense information in the prompt. Overall, the findings suggest that a well tuned multilingual encoder with the right linguistic signal can still beat larger instruction tuned models on idiom detection, especially when we have enough labeled data in at least one language.

Error Analysis

Most errors occur on borderline cases where the same idiom can be read as slightly figurative or slightly literal depending on subtle cues. Models often miss weak pragmatic signals such as sarcasm, mild exaggeration, or vague context and default to the majority sense of the idiom. This is especially visible for frequent MWEs that appear in both senses in the training data.

Cross linguistic errors show a similar pattern. The English trained XLM R sometimes over predicts the idiomatic class in Portuguese and low resource languages when the context is short or noisy. LLMs struggle with long or complex sentences and can flip the label when multiple expressions appear close together. In low resource languages the small number of examples makes predictions unstable, so a few outlier sentences can strongly affect macro F1.

Impact & Future Directions

Our project demonstrates that idiomatic understanding remains a significant challenge for modern NLP systems, even as LLMs continue to grow in size and capability. The most immediate impact of our work is the clear evidence that larger models are not inherently better at figurative literal classification. In many cases, XLM-R a smaller, structured encoder outperformed both LLaMA and Qwen, especially when supported with WSD. This finding contributes to a broader understanding in the AI community that model scale alone does not guarantee semantic reliability, particularly for culturally grounded expressions.

Another important impact of our work is its contribution to multilingual evaluation. By creating datasets for Punjabi, Bengali, and Malayalam languages that are often overlooked in NLP research , we highlight performance disparities between high resource and low resource languages. Our results reinforce the need for more inclusive datasets and evaluation methods, and they demonstrate the value of student driven linguistic contributions. Future research can build on our annotated examples or extend this methodology to additional languages and idiom categories.

The project also surfaces critical insights about model failure modes, including literal overgeneralization, hallucinated explanations. These examples can inform designers of educational tools, translation systems, and user facing AI applications, where misunderstandings of figurative meaning can have real world consequences. Our work positions idiom understanding as a practical diagnostic for semantic robustness.

Looking ahead, several promising directions emerge. One is to expand our dataset into a full multilingual benchmark with more idioms per language and varied figurative constructions such as metaphors, similes, and proverbs. Another direction is to explore retrieval augmented methods, inspired by our interview with Dr. Hauer, to anchor model predictions using definitions, examples, or sense inventories. Evaluating these systems in tasks such as machine translation, auto dubbing, or educational tools could reveal how figurative misinterpretation propagates into real-world outputs.

Finally, training lightweight or specialized models specifically designed for multiword expressions represents a longer term opportunity. With deeper collaboration across linguistics and AI, future work could produce systems that genuinely understand idiomatic meaning rather than mimicking patterns.

Overall, our case study contributes new multilingual insights, highlights important gaps in current AI systems, and lays the groundwork for more culturally inclusive research on figurative language understanding.

Conclusion

This project set out to examine whether modern AI systems truly understand idiomatic meaning or simply rely on surface-level cues. Through systematic evaluation across five languages, comparison of encoder models and LLMs, and the integration of WSD augmentation, we found clear evidence that idioms remain a difficult semantic challenge for current NLP systems. XLM-R consistently outperformed larger instruction tuned LLMs, demonstrating that structure and supervision often matter more than scale. Our multilingual results also revealed significant gaps in low resource languages, underscoring the uneven linguistic coverage of today's models.

Beyond the quantitative outcomes, the project highlighted deeper issues in AI reliability. The failure cases literal overgeneralization, hallucinated reasoning, and inconsistent predictions show that confident model outputs cannot be equated with genuine understanding. These insights have implications for translation tools, educational technologies, and any application where figurative language plays a role.

By combining established datasets with our own hand annotated examples, and by incorporating linguistic insights from our interview with Dr. Hauer, we created a replicable, meaningful framework for studying idiomatic interpretation. Our work not only contributes new multilingual observations but also opens the door for future research on retrieval-based methods, expanded idiom inventories, and improved semantic grounding.

In the end, this project demonstrates that while AI can process language fluently, true semantic understanding, especially of culturally embedded expressions remains an open and important challenge.