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 poses a persistent challenge for contemporary NLP because idioms, proverbs, and metaphors are non-compositional and highly context dependent. Even when models appear fluent, they frequently default to literal readings, offer overconfident explanations of incorrect interpretations, or comply too readily with leading prompts. These behaviours produce practical harms - misunderstanding, cultural bias, and erosion of user trust, and they also expose gaps in models' ability to select meaning from context rather than from surface form alone.

In Milestone 2 we narrow the problem to idiomaticity detection: given a sentence and a specified target expression, determine whether the usage is idiomatic or literal. We refine scope for feasibility and comparability by anchoring evaluation to established English resources that provide official splits and metrics. The technical framing is a straightforward classification setup, allowing us to study errors on carefully constructed minimal pairs where the same expression appears once idiomatically and once literally. The audience for the artifact is practitioners and students who need an interpretable, reproducible way to test whether models are choosing senses from context.

The central research question is whether a small, inference-time semantic cue can measurably improve decisions without retraining models or altering benchmark conditions. Concretely, we will compare two baselines: a supervised encoder classifier and a lightweight instruction-tuned LLM against the same models augmented with a short, neutral "sense card" presented only at decision time for the dataset's target expression. The sense card lists a literal meaning and an idiomatic meaning (and occasionally a concise distractor); the model must select the meaning that fits the sentence's context, and this choice is mapped to idiomatic vs. literal. This setup keeps inputs and scoring identical to the underlying benchmarks while probing whether explicit alternatives encourage genuine context use.

To ensure feasibility within the project timeline, we prioritize English test splits for the main quantitative results and reserve a small, clearly caveated extension to additional languages only if time and coverage permit. Success will be measured by improvements in accuracy and macro-F1 relative to both baselines, along with simple robustness checks such as option-order shuffling and neutral wording to rule out prompt leakage. The artifact we intend to deliver is a compact, reproducible evaluation package - data indices, scripts, and a short report - that clarifies where models still fail on context-dependent idioms and whether a minimal decision-time cue can reduce those characteristic mistakes.

Literature Review

Research on idiomaticity detection with encoder models has consistently treated the task as supervised classification over potentially idiomatic expressions, showing steady gains when models are trained with additional structure. Recent work demonstrates that incorporating cross-lingual “translation drift” and word-cohesion signals into BERT-style systems improves sequence-level accuracy and generalization across multiple idiom datasets, establishing strong supervised baselines for usage decisions (Yayavaram et al., 2024). These approaches motivate our inclusion of a fine-tuned encoder as a reference point, while also highlighting that better disambiguation often comes from injecting structured information.

Community benchmarks have standardized how idiom understanding is evaluated. SemEval-2022 Task 2 provides multilingual data, explicit target expressions, and official accuracy/F1 metrics, confirming that idiomatic expressions continue to challenge both monolingual and multilingual systems and enabling apples-to-apples comparison across submissions (Tayyar Madabushi et al., 2022). Building on such protocols ensures that any improvements we observe are not artifacts of custom datasets or scoring, and it lets us position results directly alongside prior reports.

The emergence of instruction-tuned LLMs shifted the conversation from whether models “know” idioms to whether they can choose the correct sense from context on deliberately difficult items. An expert-curated English test suite designed to be hard shows that conversational LLMs still make systematic errors, including false positives in clearly literal contexts and failures on adversarial minimal pairs, evidence that prompting alone is unreliable for context-sensitive disambiguation (De Luca Fornaciari et al., 2024). Complementary evaluations find that LLMs perform well on prototypical idioms but degrade when literal cues - such as motion, concrete objects, or locative phrases are present, underlining the need for targeted decision-time support rather than generic instruction prompts (Phelps et al., 2024).

Cross-language analyses further document variability by figurative type and language, with simple prompting tactics helping inconsistently and model choice having a significant effect, especially outside high-resource English. Recent comparative work across idioms and similes reports uneven performance and argues for small, model-agnostic interventions that encourage genuine context use without heavy retraining (Khoshtab et al., 2025). Publicly available idiom inventories also expand coverage beyond English; collections that aggregate Indian-language idioms and proverbs, including Punjabi and Malayalam, can seed phrase inventories and examples for later non-English probing under the same decision protocol (Tandon, 2023). Taken together, supervised encoder baselines, standardized evaluations, hard LLM test suites, multilingual variability, and accessible inventories converge on a shared conclusion: the critical open problem is not superficial familiarity with expressions but reliable, context-based sense selection.

Our project situates itself in this space by testing whether a minimal, explicit presentation of plausible meanings at decision time nudges models toward the correct choice under unchanged benchmark conditions.

Metrics of Success

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

Informational Interview

Interviewee:

Dr. Bradley Hauer, Postdoctoral Researcher in Natural Language Processing (Computational Lexical Semantics), University of Alberta.

Summary of Interview:

We spoke with **Dr. Bradley Hauer**, a postdoctoral researcher specializing in Natural Language Processing, to validate our motivation for **IdiomSense** and gain deeper insight into how large language models (LLMs) handle figurative language.

Dr. Hauer explained that while LLMs are remarkably fluent, they often *hallucinate* idiom meanings and "can't always justify why." As he put it, *"LLMs can explain anything, including nonsense. You can ask it to explain something and it will give an explanation, it just won't make any sense."* He pointed out that this confidence in incorrect answers reveals a lack of true semantic understanding.

He also noted that LLMs tend to perform well in English but struggle with other languages. *"A lot of LLMs that are really strong in English completely fall apart on relatively uncommon languages - Chinese is a good example. For complex text classification tasks, the performance drops off a cliff,"* he said.

Dr. Hauer emphasized that integrating retrieval-based or sense-aware components could significantly improve interpretability and performance, especially when dealing with idioms across multiple languages. He added that *"retrieval or external sense references could add real value"* in helping models stay grounded and accurate. His insights reinforced that idiomatic understanding remains both a relevant and technically challenging problem in modern NLP.

Reflection and Application

This interview helped us refine our direction for **IdiomSense**. We decided to focus on **inference-time disambiguation** using compact *"sense cards"* that guide model interpretation, rather than relying solely on training data. Dr.Hauer's feedback also inspired us to continue to include **multilingual idioms** in our evaluation and to experiment with **open-weight models** for better transparency.

Additionally, we plan to go beyond basic accuracy metrics and include **F1-score** and **contextual reasoning** to capture deeper understanding. Overall, his insights validated that exploring lightweight, interpretable methods for idiomatic understanding is both necessary and underexplored, giving our project a clear and meaningful direction.

Ethical, Safety, and Risk Concerns

Our project, *IdiomSense*, focuses on improving large language model (LLM) understanding of idioms using inference-time “sense cards.” Although the project does not involve personal data, several ethical and safety considerations remain relevant.

First, **cultural and linguistic bias** poses a risk. Idioms are culturally specific, and English-centric datasets such as IdioTS may underrepresent idioms from other languages or dialects. To address this, we plan to test at least one non-English language (L2) and transparently report cross-linguistic performance differences.

Second, **misinterpretation and overgeneralization** can occur when models incorrectly classify literal phrases as idiomatic. Because figurative meaning is highly context-dependent, we evaluate detection on minimal pairs to ensure that decisions reflect genuine contextual reasoning rather than memorization.

Third, **explainability and dataset ethics** are important. Sense cards inherently improve transparency by exposing the cues influencing model predictions. All resources used (WordNet, BabelNet, IdiomKB) are open-access and will be properly credited according to their licenses.

Finally, we acknowledge potential **bias or misuse** in model outputs. We will responsibly report both successful and failed cases to present a balanced, transparent account of our system’s limitations and ethical implications.

Team Reflection

Team Member	Specific Tasks Completed So Far
Mohammad Shahriar Hossain	Proposed and explored different methods for identifying idioms and proverbs in language datasets, contributed to shaping the project's core idea and created slides for presentations
Donna Mathew	Found and analyzed relevant research papers; helped connect theoretical concepts to our project direction and created slides for presentations
Prabal Mehra	Helped organize meetings, developed interview questions, and supported decision making discussions.Assisted with analyzing research papers, and created slides for presentations

Part 2: Team Communication and Collaboration

- Communication has been excellent, with responses to messages typically under 4-5 hours.
- Work is evenly divided, ensuring meaningful contributions from all members.
- During midterm periods, progress slowed slightly, but team members were supportive in picking up slack so everyone could focus on both midterms and the project.
- All team members were open to asking for help which strengthened collaboration and built a supportive environment.
- Overall, the team combined accountability with a positive, cooperative dynamic .

Plan

Date	Milestone	Description
Oct 30	1	**Data ready for baselines** • Collected and formatted items from IdioTS, SemEval-2022 Task 2, and the Kaggle multilingual idioms list. • Unified JSONL/CSV schema: {sentence, target_expression, label, split}. • Official splits respected for SemEval/IdioTS; a tiny held-out slice created for Kaggle items. **Responsible Member(s):** Prabal Mehra (lead data ingestion & schema), Donna Mathew (split validation & QA), Mohammad Shahriar Hossain (dedup/normalization scripts)
Nov 6	2	**Baselines completed on English** • BERT idiomaticity detector trained/evaluated on official EN splits; predictions saved. • Lightweight open LLM (e.g., Gemma/LLaMA) run with a short decision prompt; outputs + logs saved. • First metrics computed: Accuracy, F1(idiomatic), Macro-F1; quick sanity error list. **Responsible Member(s):** Donna Mathew (lead BERT training/eval), Mohammad Shahriar Hossain (LLM prompt runs & logging), Prabal Mehra (metric scripts & sanity error list)
Nov 13	3	**Sense-card runs + ablations** • Minimal sense-card builder implemented (≤30 tokens per meaning; literal + idiomatic, optional distractor). • Re-run BERT and LLM with sense cards on the same test items. • Ablations: no-card vs card; option order shuffle; neutral wording. • Updated metrics + robustness notes. **Responsible Member(s):** Prabal Mehra (sense-card builder & coverage logs), Donna Mathew (BERT+card runs; ablations for BERT), Mohammad Shahriar Hossain (LLM+card runs; ablations for LLM; compile robustness notes)
Nov 20	4	**Wrap-up** • Final result tables (baseline vs sense-card), 95% CI (bootstrap) and McNemar where feasible. • Compact error analysis with 8–12 illustrative examples (false-idiomization vs false-literal). • Optional small multilingual probe (BN/PA/ML) using Kaggle items; qualitative notes. • Reproducibility: scripts + README/Makefile; brief report and slides complete. **Responsible Member(s):** Prabal Mehra (final tables; significance tests; repo Makefile/README), Donna Mathew (error analysis & examples; brief report write-up), Mohammad Shahriar Hossain (multilingual probe slice; slide deck & packaging)

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