MINIMAL-CONTEXT STRESS TESTS OF FIGURATIVE REASONING IN LARGE LANGUAGE MODELS

Yawen Guo (Student)
School of Computer Science
McGill University
Montréal, QC H3A 2A7
{yawen.guo2}@mail.mcgill.ca

Balaji Ramesh (Student)
School of Computer Science
McGill University
Montréal, QC H3A 2A7
{balaji.ramesh}@mail.mcgill.ca

Zarine Ardekani-Djoneidi (Student)
School of Computer Science
McGill University
Montréal, QC H3A 2A7
{zarine.ardekani-djoneidi}@mail.mcgill.ca

Verna Dankers (Mentor)
School of Computer Science
Mila and McGill University
Montréal, QC H3A 2A7
{verna.dankers}@mila.quebec

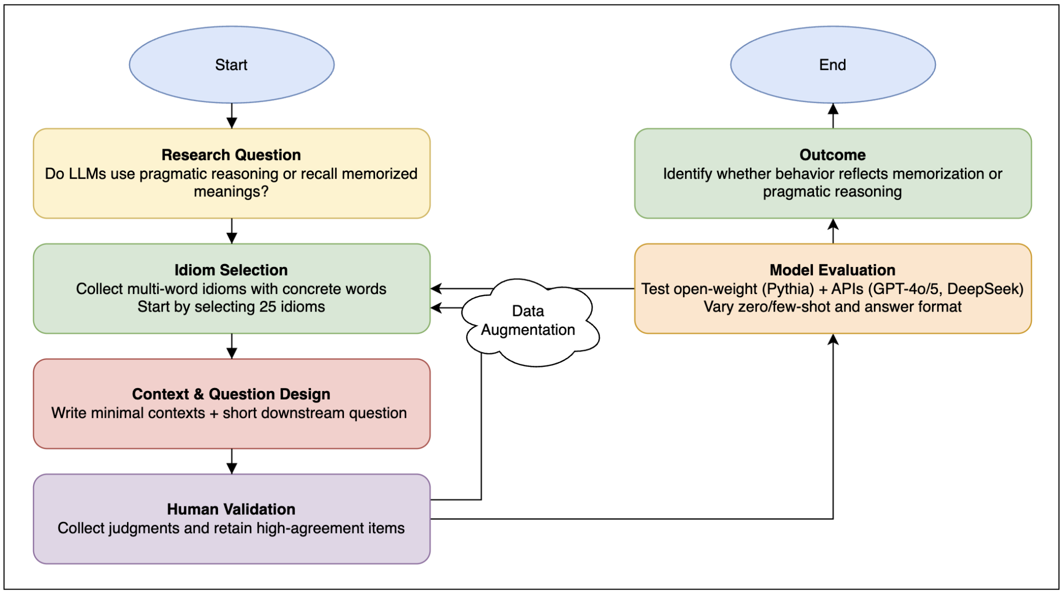
Start
Research Question
Do LLMs use pragmatic reasoning or recall memorized meanings?
Idiom Selection
Collect multi-word idioms with concrete words
Start by selecting 25 idioms
Context & Question Design
Write minimal contexts + short downstream question
Human Validation
Collect judgments and retain high-agreement items
Data Augmentation
Model Evaluation
Test open-weight (Pythia) + APIs (GPT-4o/5, DeepSeek)
Vary zero/few-shot and answer format
Outcome
Identify whether behavior reflects memorization or pragmatic reasoning
End

Figure 1: Proposed methodology. (diagrams.net, 2025)

Motivation Idioms are hard for LLMs because correct interpretation depends on context rather than surface form (Miletić & Walde, 2024; Flor et al., 2025). Recent work often frames idiomaticity as binary classification (figurative vs. literal), sometimes as sequence labeling or token-level (idiom span detection), and builds large resources such as MAGPIE and ID10M to support evaluation (Haagsma et al., 2020; Tedeschi et al., 2022). Yet these settings can reward superficial cues or memorized glosses (Mi et al., 2025; Haviv et al., 2023). Surveys of multiword expressions likewise report mixed generalization and inconsistent context use (Miletić & Walde, 2024; Flor et al., 2025). We target a specific gap: minimal-cue, human-validated probes that make a literal reading viable and then require a simple factual judgment (e.g., "dead or alive?"). Our design stress-tests pragmatic reasoning under controlled literal plausibility, not just label choice, and directly motivates the method described next.

Research Question and Method Our main question is whether today's LLMs genuinely use pragmatic reasoning to interpret idioms when a small real-world cue makes a literal reading viable, or whether they simply recall memorized meanings of the idiom. To test this, we first gather data, then write minimal contexts and a short question, then human-validate and keep only items that humans correctly answer. Afterwards, we will evaluate the performance of the models. More details appear in the Experimental Design section. Our novelty is that our human-validated, minimal-context design forces a concrete inference, cleanly separating recall from pragmatic reasoning.

Hypothesis/Expected Results Prior research shows that LLMs struggle when surface shortcuts are removed (Mi et al., 2025), so we expect lower accuracy here than on standard idiom benchmarks. Because some corporas show a strong figurative skew — for example, MAGPIE reports around 70% figurative vs. 30% literal occurrences — we are expecting LLMs trained on such data to be biased toward predicting the figurative sense. (Haagsma et al., 2020)

Experimental Design We will do inference only (no training). For models, we will use open-weight scaling families (e.g., Pythia at several sizes) and large API models (e.g., GPT-4o, DeepSeek). For data, we will curate multi-word idioms (≥ 3 words) from reputable sources (e.g., MAGPIE) and start with 25 whose concrete word offers a clear literal cue and mark the cue word for each idiom while discarding idioms that lack a straightforward literal cue, and write minimal contexts that make the literal reading plausible plus a short downstream question whose correct answer depends on the intended sense. We will run human studies and keep only items that humans correctly answer; anything ambiguous is discarded. To evaluate model performance, we vary zero vs few-shot prompts (with and without in-context examples – i.e., zero-shot has no examples; few-shot includes 2–3 in-context $Q \rightarrow A$ examples) and one-word only vs free-text answers, and test for anti-scaling trends across model sizes. The main metric is accuracy. Data collection and computational experiment will be performed iteratively and in parallel. We plan to iteratively expand our dataset beyond the initial 25 entries.

For example, an experiment setup for one without in-context example, one-word only answer prompt: "Consider the sentence 'Barney's doodles were found on the same page as his recipes.' Is this claim about documents or about people?" In this case, the idiom is "on the same page" that we found from MAGPIE (Haagsma et al., 2020), which if using figuratively, means "(of two or more people) in agreement. US". (Ayto, 2020) We select "page" as the cue word.

Project Timeline and Roles Everyone will contribute across all stages of this project. Balaji will coordinate data augmentation while Yawen and Zarine will coordinate data analysis. Our mentor will approve the idiom list and prompts, review the human-study protocol and intermediate results, provide weekly guidance on analyses, and approve the final submission.

- **Oct 20 – 26:** Collect data and create contexts and questions.
- **Oct 27 – Nov 2:** Conduct human studies and continue working on the dataset, prompts and evaluation scripts.
- **Nov 3 – 23:** Run inference on models and compute metrics. Augment dataset as needed.
- **Nov 24 – 30:** Complete final analyses, write the report, review with our mentor, and submit.

REFERENCES

John Ayto. *Oxford Dictionary of Idioms*. Oxford Quick Reference. Oxford University Press, Oxford, 4 edition, 2020. ISBN 9780198845621. doi: 10.1093/acref/9780198845621.001.0001. URL https://www.oxfordreference.com/display/10.1093/acref/9780198845621.001.0001/acref-9780198845621.

diagrams.net. diagrams.net (app.diagrams.net), 2025. URL https://app.diagrams.net/. Accessed October 20, 2025.

Michael Flor, Xinyi Liu, and Anna Feldman. A Survey of Idiom Datasets for Psycholinguistic and Computational Research, August 2025. URL http://arxiv.org/abs/2508.11828. arXiv:2508.11828 [cs].

Hessel Haagsma, Johan Bos, and Malvina Nissim. MAGPIE: A Large Corpus of Potentially Idiomatic Expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 279–287, Marseille, France, 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.35/.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. Understanding Transformer Memorization Recall Through Idioms, February 2023. URL http://arxiv.org/abs/2210.03588. arXiv:2210.03588 [cs].

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context, July 2025. URL http://arxiv.org/abs/2410.16069. arXiv:2410.16069 [cs].

Filip Miletić and Sabine Schulte Im Walde. Semantics of Multiword Expressions in Transformer-Based Models: A Survey. *Transactions of the Association for Computational Linguistics*, 12:593–612, April 2024. ISSN 2307-387X. doi: 10.1162/tacl_a_00657. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00657/120831/Semantics-of-Multiword-Expressions-in-Transformer.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. ID10M: Idiom Identification in 10 Languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2715–2726, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.208. URL https://aclanthology.org/2022.findings-naacl.208.