Effects of Source Language on Idiom Interpretation

Anjali Ragupathi

March 21, 2024

1 Introduction

1.1 Background and Motivation

We begin this work by defining some key terms. An idiom is a non-compositional, conventionally used phrase whose meaning cannot be easily deduced from a combination of the meanings of its component words. A metaphor, by contrast,

1.2 Hypotheses and Assumptions

1.3 Links to GitHub and OSF Preregistration

The GitHub repository for this project can be found at this link.
The experiments conducted for this project have been preregistered with the Open Science Foundation at this link.

2 Experiment

2.1 Prior work

This project is part of a larger thesis, the objective of which is to determine if the idiom interpretation capabilities of large language models are comparable to those of humans and to what extent an underlying conceptual metaphor drives the predictions. Large language models (LLMs) are deep neural networks with billions of parameters that are trained to predict the next token or a masked token by learning statistical co-occurrences between words. Since this is done over many hours with large datasets and high compute, LLMs can ”learn” to approximate (or model) the language they are trained on.

Post-training, LLMs are often fine-tuned on more specific downstream tasks like question answering or machine translation. A more recent class of LLMs including ChatGPT and Gemini includes ”instruction-tuned” models where the inputs during the fine-tuning process are phrased as instructions or prompts. This allows for more natural interactions between humans and the models at inference time and can encompass most of the previously defined downstream tasks with additional strategies like chain-of-thought prompting (allowing LLMs to reason from the input to the predicted output step-by-step)(Wei et al., 2023) or few-shot prompting (which allows the user to provide examples of input-output pairs at inference which the model can use for improved answer quality).

For the task at hand - idiom interpretation, we compiled a dataset of 80 non-English idioms translated literally into English, the process of which is described in Section 2.2.1. The choice of translating the idioms into English was made because English could act as an anchor language; this meant that the LLM would be able to ”see” these idioms from the point of view of a primarily English-speaking listener. Additionally, English was the more suitable option as most commercial LLMs are trained on English data and it would be easier to integrate English versions of these idioms into the prompting and analysis pipeline. Finally, an important reason for this choice was that we wanted to test the extent to which the surface features of the idiom itself (i.e., constituent words and their types, associated sentiment, etc.) influenced idiom interpretation - not necessarily the ability to recall idioms (an effect of familiarity) or the influence of contextual cues from surrounding dialogue (which would mean that the listener would rely heavily on context and not on the idiomatic phrase itself). These factors motivated the choice of language and inclusion of only the idiom phrase without context of usage in conversation or larger bodies of text.

We used three different large language models for this test: HuggingChat (the default model under the hood is Llama-70B), Lllama-13B, and PaLM-2 Text-Bison-001). HuggingChat is a free LLM provided by HuggingFace and we used an unofficial open-source Python API to interface with it (API Link). PaLM-2 is a predecessor of Gemini provided by Google and Llama is provided by Meta; both were trained on diverse, multilingual datasets and evaluated for potential harms.

Eight variants of the same prompt were constructed, including ”Interpret the figurative meaning of this idiom”, ”Retrieve the figurative meaning of this idiom”, ”This is the literal translation of an idiom: $\langle IDIOM \rangle$. What is its figurative meaning?”, as well as versions which asked the model to provide a step-by-step reasoning for its prediction, gave the model some room for experimentation by asking what the model ”thought” the figurative meaning was (models were less likely to hedge their answers), and asking the model to make the prediction ignoring the language of origin (to potentially eliminate any shortcuts the model could take by just retrieving the meaning and not coming up with plausible answers). The step-by-step prompt results were subsequently excluded from further analysis in order to standardize result format and make subsequent analyses easier, but they were retained for potential future analyses.

Each model ran inference on the entire dataset of 80 idioms thrice to account for variation in response and to make the resulting evaluation strategies more robust. Some of these free-form responses were then used as confound options for the current project.

2.2 Methods

2.2.1 Materials

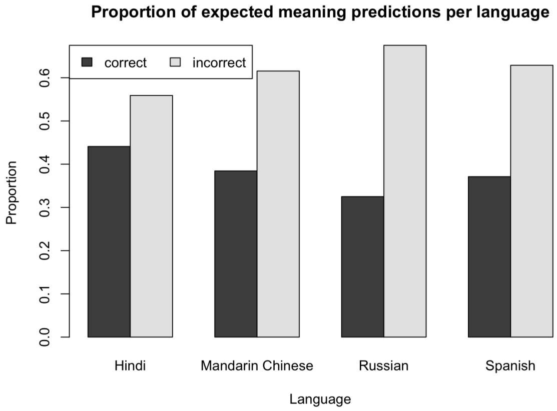
A list of 80 idioms was prepared by collecting the literal translations (into English) and corresponding figurative meanings of different idioms from Russian, Spanish, Hindi, and Mandarin Chinese from language-learning websites and publicly available digital school textbooks. 20 idioms were included from each language for balance.

For each idiom, in addition to the expected figurative meaning, three confounds were prepared. In reference to the outputs of the previous experiment (see Section 2.1), confounds were chosen from among the plausible responses produced by the large language models. A confound here is defined as a phrase that conveys a plausible potential meaning as either produced by a large language model or one that relates to a relevant underlying conceptual metaphor. For cases where large language models produced largely consistent results, the latter method of defining confounds was used and the experimenter manually created entries for those confounds.

In addition to the 80 cross-lingual idioms, 4 English idioms - namely, ”once in a blue moon”, ”to face the music”, ”to break the ice”, and ”to feel under the weather” - were selected as attention check idioms. They served as a means of standardizing participants' understanding of what an idiom was and if they were paying attention during the experiment instead of randomly selecting options.

2.2.2 Participants

A pilot study with 10 participants was run to determine estimated completion time and any difficulties with conceptual and technical components of the experiment. Participant recruitment was done on Prolific and restricted to the US, UK, EU, and India. Based on the results of this experiment, the initial instructions were modified to include an example of an idiom, i.e. ”kick the bucket”, so that participants would be primed towards the format of the experiment. The estimated completion time was also reduced to 12 minutes from the originally expected 15 minutes after the pilot study.



The main study recruited 100 participants on Prolific and included participants across the US, UK, EU, India, Singapore, Switzerland, and Mexico, to encourage a diversity of participants to take part in the study. The main study was released in two batches (50 participants each) at two different time slots to gather data from participants in different global time-zones. Nevertheless, the primary demographic of the final participant pool was from residents of the United Kingdom and the United States.

2.2.3 Procedure

Each participant was shown 20 idioms - 4 from each target language and 4 attention check idioms - randomly shuffled, and presented in a sequence. Each idiom was presented along with the 4 options (the expected figurative meaning and the three confounds) which were also shuffled. Participants saw one idiom at a time and were required to select an option before moving on, with no option to go back and modify their response. Towards the end of the experiment, participants were asked to fill a survey with demographic information like age, gender, level of education, and languages they were natively proficient in (that is, language(s) they grew up speaking or possessed the ability to fluently write, read, and speak in to a native level).

2.3 Results

2.3.1 Outcomes of the Experiment

For the purposes of this study, we define ”correct” to be a situation where the participant predicted the expected meaning (corresponding figurative meaning of the translated idiom in its original language) and ”incorrect” to be a situation where they predicted a different meaning. Naturally, these are simplifications to the actual nature of such an experiment and we do not aim to test if people can predict a correct meaning as these idioms simply do not exist in English; it would be an invalid assumption to make that people can predict the correct meaning of a non-existent idiom.

In Figure

2.3.2 Mixed-Effects Logistic Regression Model

References

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.