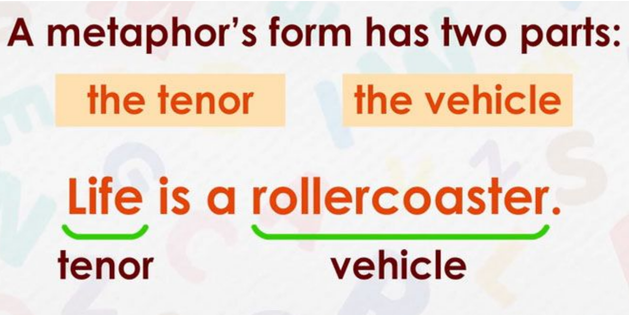
Final Project

William Palmer, Sivan Almogy, Avi Kabra

Motivation

- Metaphors are a hallmark of human cognition
- Although humans intuitively grasp these relationships, machines must overcome significant barriers in representation and reasoning to achieve a similar level of understanding
- **Do LLMs possess the requisite content knowledge of attribute-vehicle relationships independent of metaphorical structure?**



Motivation (cont’d)

- Machines often struggle with both components due to the abstract, nonliteral nature of these comparisons.
- While models perform well in surface-level tasks, they often fail to generalize the object-adjective relationships accurately

- Do these failures stem from a lack of knowledge or limitations in application?

Cultural and Conceptual Knowledge

- Vital for understanding metaphors because these comparisons often emerge from shared experiences and collective perception
  - Second-nature for humans
  - Machines: lack of real-world experience or embodied cognition
- LLMs must infer these relationships solely from textual data, which may lack the context or depth required to build such associations reliably
- A single vehicle can convey multiple attributes depending on usage
  - ie. “iron will” vs. “iron rule”
- Co-occurrence of words vs. deeper semantic relationships

Existing Literature

Veale et al. (2008)

- Utilized lexical databases like WordNet to identify plausible mappings between vehicles and targets
- Approach relied heavily on pre-structured knowledge, limiting its generalizability to novel metaphors

Tsvetkov et al. (2014)

- Leveraged multilingual data to train computational models capable of identifying metaphorical expressions across languages
- Performance declined when encountering metaphorical expressions that required cultural grounding

Existing Literature (Semantics)

Shutova et al. (2013): LLMs struggled with interpreting metaphors requiring deeper conceptual knowledge or cultural context

Bizzoni et al. (2017): LLMs excel at identifying surface-level similarities between words, but they often fail to encode the richer, non-literal associations necessary for metaphor interpretation

Dankers et al. (2020): Argued that LLMs lack the embodied knowledge humans use to disambiguate figurative expressions

Takeaways

- Studies highlight significant progress and persistent limitations in computational metaphor processing
- While previous research has explored metaphor application, far less attention has been given to evaluating whether LLMs possess the underlying content knowledge of attribute-vehicle relationships
  - Failures in metaphor comprehension stem from knowledge deficits vs. from difficulties in applying existing knowledge within figurative structures

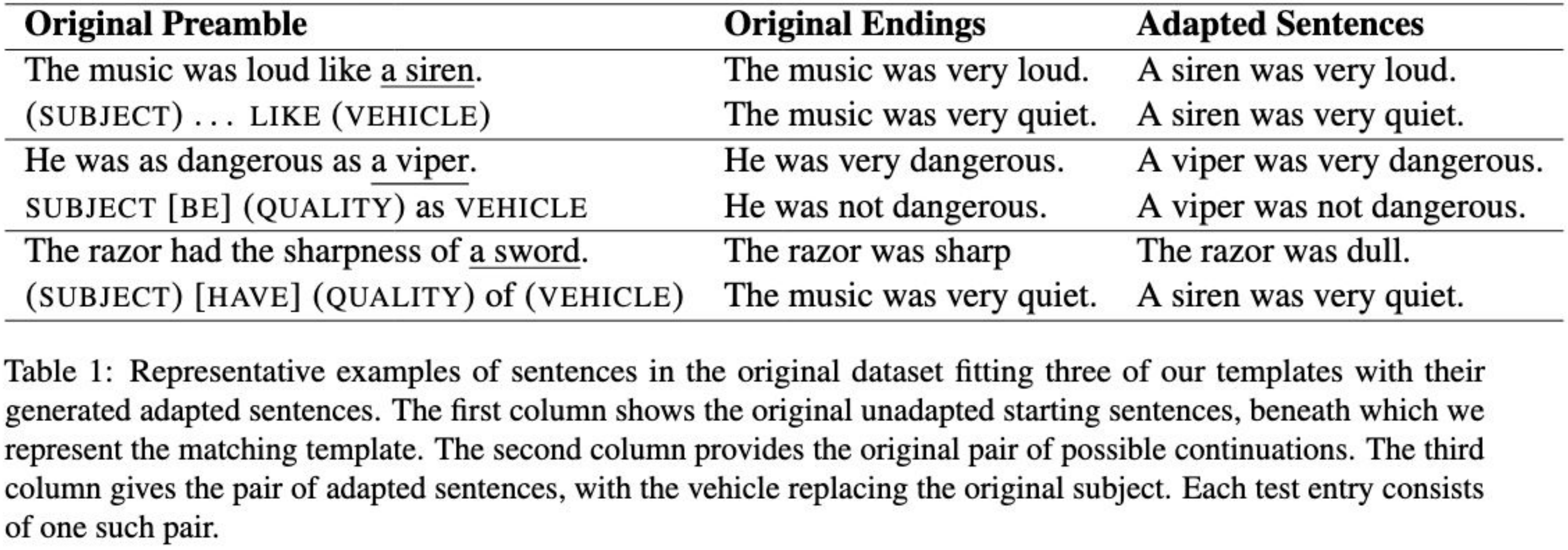
Do LLMs possess knowledge of **attribute-vehicle relationships** outside figurative structures?

Our Approach

- Metaphor interpretation dataset constructed by Liu et al. (2022)
  - Predicting relative likelihoods of paired sentences given a metaphor preamble
  - LLMs (GPT2, GPT3, BERT) performed well below human accuracy
- We aimed to isolate the needed background knowledge
  - Poor comprehension could be explained by lacking relevant information
- Generated an adapted version of Liu et al.'s devset sentences

Adapting the Dataset

- Created 5 starting sentence patterns, adapted ~800/1000 entries
- For each possible ending, subject replaced with metaphor vehicle



Original Preamble	Original Endings	Adapted Sentences
The music was loud like a siren.	The music was very loud.	A siren was very loud.
(SUBJECT) . . . LIKE (VEHICLE)	The music was very quiet.	A siren was very quiet.
He was as dangerous as a viper.	He was very dangerous.	A viper was very dangerous.
SUBJECT [BE] (QUALITY) as VEHICLE	He was not dangerous.	A viper was not dangerous.
The razor had the sharpness of a sword.	The razor was sharp	The razor was dull.
(SUBJECT) [HAVE] (QUALITY) of (VEHICLE)	The music was very quiet.	A siren was very quiet.

Table 1: Representative examples of sentences in the original dataset fitting three of our templates with their generated adapted sentences. The first column shows the original unadapted starting sentences, beneath which we represent the matching template. The second column provides the original pair of possible continuations. The third column gives the pair of adapted sentences, with the vehicle replacing the original subject. Each test entry consists of one such pair.

Evaluating performance

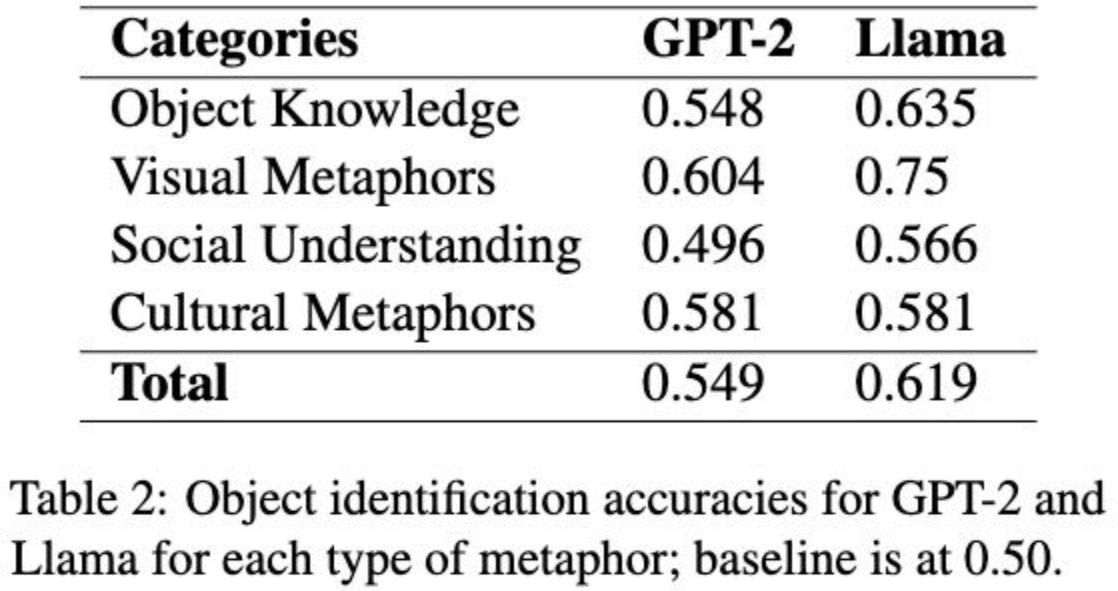
- Given modified sentences for the same object, which is more likely?

  E.g. “A siren was very loud” vs. “A siren was very quiet”
- “Predict” ending with higher sum of log probabilities
  - Difference between paired sentence lengths has SD ~1.5

- Also did comparison for same ending modified with each object

  E.g. “A siren was very loud.” vs. “A whisper was very loud.”
- Ending with higher sum of log probabilities taken *after* the subject
  - Conditional probabilities: P(“was very loud” | “A siren”) and P(“was very loud” | “A whisper”)
  - No difference in length

Results for Object Identification

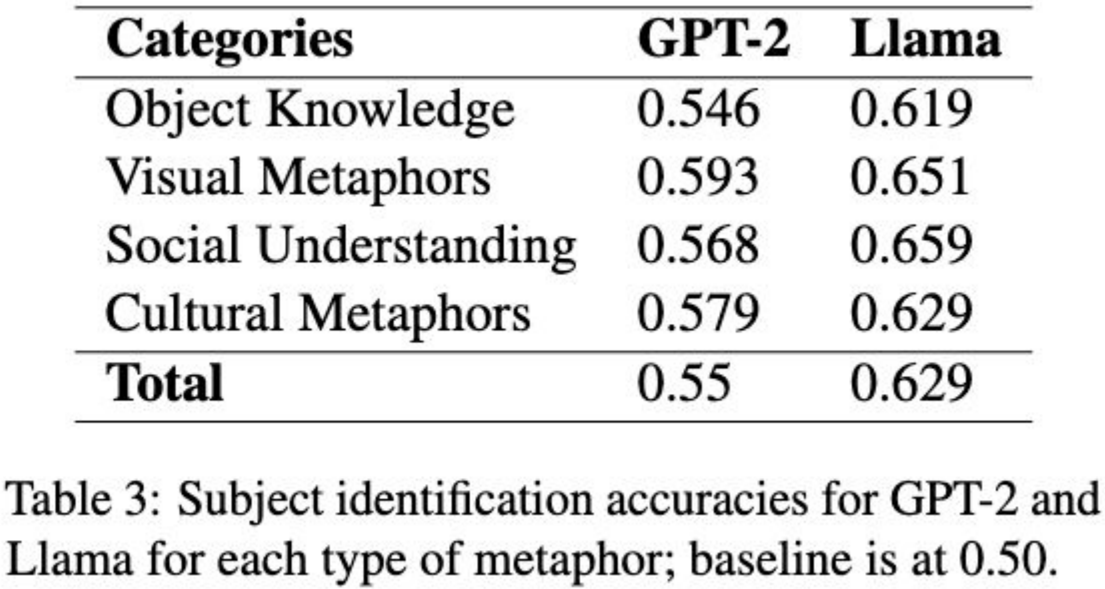


Categories	GPT-2	Llama
Object Knowledge	0.548	0.635
Visual Metaphors	0.604	0.75
Social Understanding	0.496	0.566
Cultural Metaphors	0.581	0.581
Total	0.549	0.619

Table 2: Object identification accuracies for GPT-2 and Llama for each type of metaphor; baseline is at 0.50.

GPT2 had 53.93% accuracy on Liu et al.’s original evaluation.

Results for Subject Identification



Categories	**GPT-2**	**Llama**
Object Knowledge	0.546	0.619
Visual Metaphors	0.593	0.651
Social Understanding	0.568	0.659
Cultural Metaphors	0.579	0.629
Total	0.55	0.629

Table 3: Subject identification accuracies for GPT-2 and Llama for each type of metaphor; baseline is at 0.50.

Limitations and Next Steps

Issues with dataset generation

- Subject-vehicle replacement does not always capture correct information
- In the second case, there is also a incorrect swapping / identification of subject
  - Still, no clear way to adapt follow up sentences to isolate needed information
  - "He was very certain of concrete” also a poor adaptation
- Ideally, new sentences would be created manually

Original Preamble	Original Endings	Adapted Sentences
Shopping for groceries is a scavenger hunt with a list created by a lunatic	Shopping for groceries is a fun, rewarding chore Shopping for groceries is a crazy, nearly impossible chore	A scavenger hunt with a list created by a lunatic is a fun, rewarding chore A scavenger hunt with a list created by a lunatic is a crazy, nearly impossible chore
His opinions were as firm as (concrete / a cotton ball)	He was very certain of his opinion He was very uncertain of his opinion	Concrete was very certain of his opinion Concrete was very uncertain of his opinion

Future steps

- Fine-tuning models or modifying prompts to provide requisite information
  - When the knowledge is given, how well can LLMs apply it?
- Directly comparing accuracy on each entry in Liu et al.’s evaluation
  - Could reproduce their results, check if simile interpretation accuracies correspond with ours
- Improved knowledge evaluation data