How to evaluate large language models: Insights from linguistics

Tom McCoy,
tom.mccoy@yale.edu
Yale Department of Linguistics

Zhenghao Herbert Zhou
herbert.zhou@yale.edu
Yale Department of Linguistics

1 Overview

- AI users need ways to evaluate the capabilities of AI systems
- Much of the recent progress in AI comes from systems based on language (large language models)
- Given the centrality of language in these models, perhaps the field of linguistics can help us understand them!
- **Main claim:** Methods used in linguistics to understand the human mind can also be used to evaluate language-based AI systems
- **Approach discussed in this poster:** the minimal pair paradigm

2 The Minimal Pair Paradigm

- Minimal pair: A pair of examples, one of which is well-formed and one of which is ill-formed
- Used in linguistics to illustrate the rules of grammar

Example minimal pair (* means ill-formed)
1 a. The dog is barking
b. *The dog are barking
Illustrates the rule of subject-verb agreement

- Can also be used to test what types of knowledge and skills are captured by large language models
  - Language models assign probabilities to strings of text
  - To test a language model with a minimal pair: See if it assigns higher probability to the well-formed example than the ill-formed one

3 Example: Subject-Verb Agreement

Initial test: Simple examples

- Evaluate on minimal pairs like 1a vs. 1b
- Language models score 94% to 100% (Marvin & Linzen 2018)
- Evidence that they have captured subject-verb agreement!

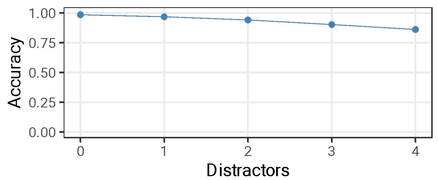
Further tests: Considering confounds

- It's important to control for confounds!
- **Follow-up 1:** What if models just have a verb agree with the closest noun (rather than the subject)?

Test: Add "distractors" (underlined)

2 a. The lawyer by the doctors is laughing
b. *The lawyer by the doctors are laughing

Result: Model continues to succeed even with distractors (Dyer et al. 2018)

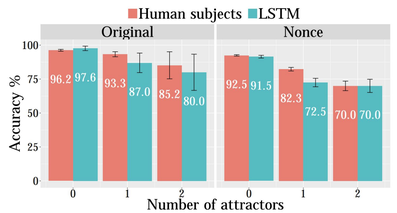


- **Follow-up 2:** What if the model just memorizes subject-verb pairs?

Test: Use nonsensical subject-verbs pairs that the model would not have seen

3 a. The lamp by the happiness smiles
b. *The lamps by the happiness smiles

Result: Only small degradation in accuracy (Gulordava et al. 2018)



4 Beyond Syntax: Assessing Meaning

- We can also use minimal pairs to test whether language models capture properties of sentence meaning!
- Examples from Zhu, Zhou, Charlow, & Frank (2025) and Zhu & Frank (2024); # means grammatical but contextually inappropriate

4 a. A farmer worked in his field. He dreamt of the harvest.
b. #Every farmer worked in his field. He dreamt of the harvest.

5 a. John owns a dog. The dog is cute.
b. #John doesn't own a dog. The dog is cute.

5 Beyond Language

- We can extend the use of minimal pairs even farther to evaluate non-linguistic domains!
- **Example: World knowledge**

6 a. The capital of Pennsylvania is Harrisburg.
b. #The capital of Pennsylvania is Philadelphia.

- **Example: Logical reasoning**

7 a. Socrates is a man. All men are mortal. Thus, Socrates is mortal.
b. #Socrates is mortal. All men are mortal. Thus, Socrates is a man.

6 Conclusion

- In the minimal pair paradigm, an AI system is evaluated by testing whether it favors the correct option in a pair of choices
- This paradigm can be used to create tests for a wide variety of abilities, both within language and beyond
- A recommended best practice is to use multiple conditions that control for confounds
- **High-level takeaway:** Techniques from linguistics can be used to analyze language-based AI systems