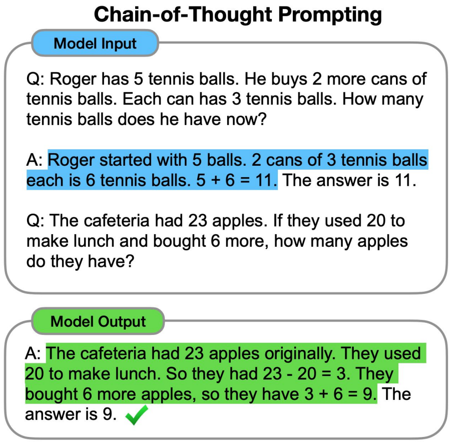
modeLing:
A Novel Dataset for Testing LLM Linguistic Reasoning

Nathan A. Chi, Teodor Malchev, Riley Kong, Ryan A. Chi, Lucas Huang, Ethan A. Chi, R. Thomas McCoy, Dragomir Radev

SIGTYP 2024

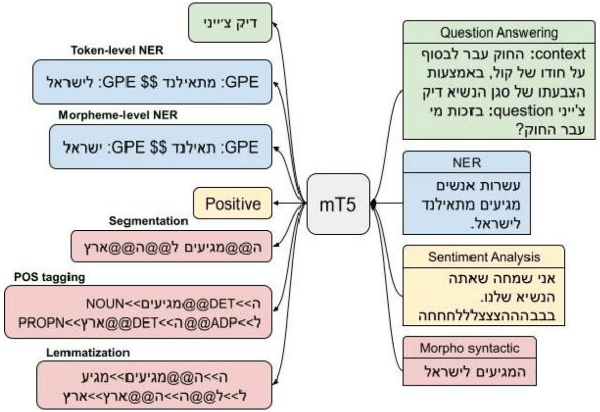
Large language models (LLMs) are good at...

reasoning



(Wei et al., 2023)

multilinguality



(Xue et al., 2020; Eyal et al. 2022)

but evaluating their intersection is tricky.

Why? **Language contamination** (Blevins and Zettlemoyer, 2022).

Overview

- We introduce **ModeLing**, a dataset that uses carefully-designed language puzzles to test **few-shot multilingual reasoning.**
- LLMs perform well on some categories in ModeLing, providing evidence that they have some few-shot multilingual reasoning capabilities
- However, there is ample room for improvement: on harder categories, performance remains poor, and models are far from perfect even on easy categories.
- These results cannot be explained by language contamination.

Rosetta stone **puzzles**

(Bozhanov and Derzhanski, 2013)

wó ùrò kàná sɔ́ɣɔ́rɔ́jɛ̀w là:
You have already unlocked his new house, haven’t you?
ójú kù^{n} námár^{n}átìm sábù ìjù téré ɛ́:tìm
I took my foot off the road because I saw a fast dog.
nìnìw^{n}é ùrò pɛ̌y^{n} náŋár^{n}átɔ́ɣɔ̀
A cat remembers an old house.
ìjú bé∴ nìnìw^{n}è těy^{n} bé∴ sǎy ànà dìgɛ́tɔ́ɣɔ̀w
You follow only dogs and small cats in the village.

- Small parallel corpus in a target language not previously known to the solver
- Corpus is chosen to uniquely specify a single most reasonable underlying set of rules

---

These puzzles originate from the **International Linguistics Olympiad** (IOL) and related secondary school competitions!

ModeLing

- Previous Rosetta Stone dataset (PuzzLing; Şahin et al., 2020) reuse problems written for Linguistics olympiads, thus raising the specter of data leakage.
- ModeLing consists entirely of **newly written questions** written specifically for this work.
- We demonstrate that popular LLMs do not display data leakage on ModeLing.

We contribute 272 Rosetta Stone questions

covering a variety of 19 less attested languages

Figure 8: The 19 distinct languages included in the MODELING benchmark. Note that some languages have more than one problem.

Anatomy of a ModeLing problem

Evidence

Here are some phrases in Ayutla Mixe:
Ëjts nexp. → I see.
Mejts mtunp. → You work.
Juan yë'ë yexyejtpy. → Juan watches him.
Yë'ë yë' uk yexpy. → He sees the dog.
Ëjts yë' maxu'unk nexyejtpy. → I watch the baby.

⚠️ Removing this section leads to 0% LLM performance, showing lack of data leakage on current LLMs.

Questions

Yë' maxu'unk yexp. → The baby sees.

The baby watches the dog. →
Yë' maxu'unk yë' uk yexyejtpy.

We ask each question separately, without the context of the other questions.

Problem Types

Noun / Adjective

Determine **relative ordering** of nouns and adjectives.

Word Order

Determine **relative ordering** of subject (S), verb (V), object (O).

Possession

Reason about **possessive morphology.**

Semantics

Align foreign semantic compounds to English translations.

Problem Types

Nominal clause order

Requires solvers to determine the relative ordering of nouns/adjectives

Bangime
tãwa nundi → "five beds"
kurɛ tiri → "one dog"
ko kiye → "seven houses"
mpa tar → "three friends"
ko tar → "three houses"
yaamɛ yinu → "two children

How to solve:
Solvers must deduce that Bangime places the modifier after the noun (*tar* "three" appears twice, both in the postnominal position.)

S/V/O order

Requires solvers to determine the ordering of subject/verb/object in a clause.

Engenni
abhwa dhi → "The dog eats."
abhwa mise → "The dog sleeps."
afeni bidha → "The bird walks."
afeni fyani → "The bird flies."
bhu dhi → "You eat."
eni dhi → "We eat."
mi bidha → "I walk."

How to solve:
Solvers must deduce that S comes before V in Engenni ("afeni", "dhi").

Problem Types

Possession

Requires solvers to determine the way possession marking works

Dogon
sáydù ìlò → "Seydou's house"
àlá-ɔ̀ŋù-nú nènù → "the village chief's dog"
í ílò → "our house"
ú nénù → "your dog"

How to solve:
Solvers must determine that 1) possessor appears before possessed 2) the tone of the first syllable changes (tone sandhi) to the tone of the last syllable.

Semantic Matching

Use cross-cultural reasoning to align foreign semantic compounds to English translations.

Kutenai
cmakwumnana → "The dog eats."
itɨcmakni → "(it) is not strong."
ɨitqatni → "(it) does not have a tail."
maknana → "little bone"
qatnana → "little tail"

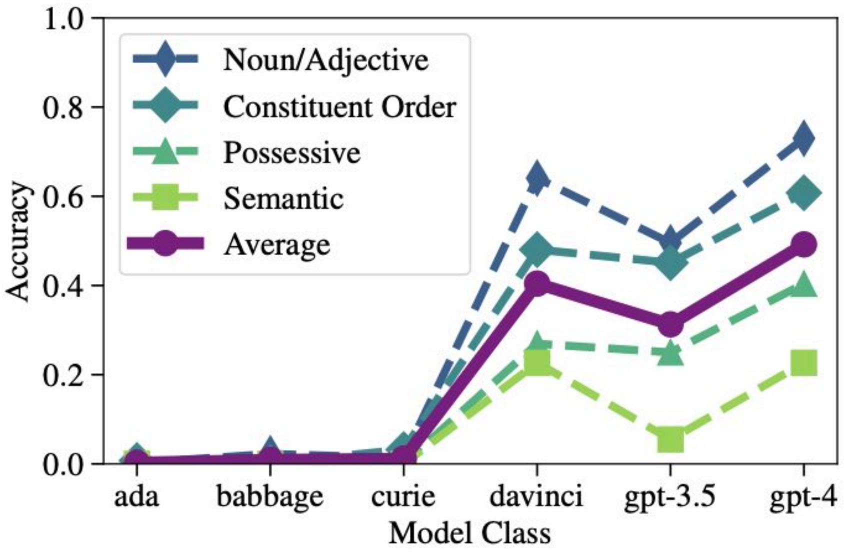
How to solve:
Solvers must perform significant semantic/morphological reasoning (e.g. -*nana* DIM, -*ni* "it does not have").

Experiments

- We evaluated all problems on **GPT-3, GPT-3.5,** and **GPT-4** as of August 13, 2023, using a number of prompting and Chain-of-Thought methods.
- We evaluated on exact match accuracy because of difficulties in using BLEU to distinguish morphological differences.

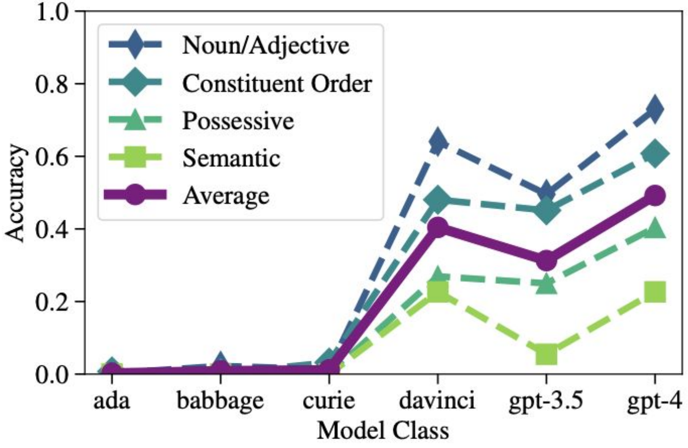
Positive Results

LLMs do pretty well! Average solve % exceeds 50% on GPT-4, and 45% on GPT-3.

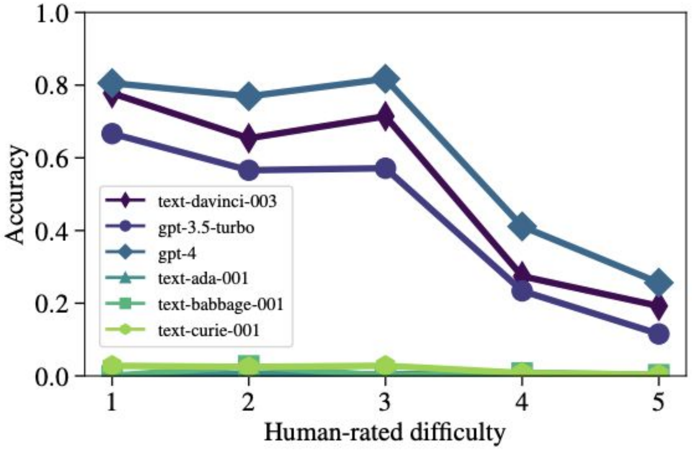


Areas for Improvement

LLMs have difficulty solving **semantic** and **possessive** problems (more complex morphology)



LLMs struggle with the **same types of questions** that humans find difficult!



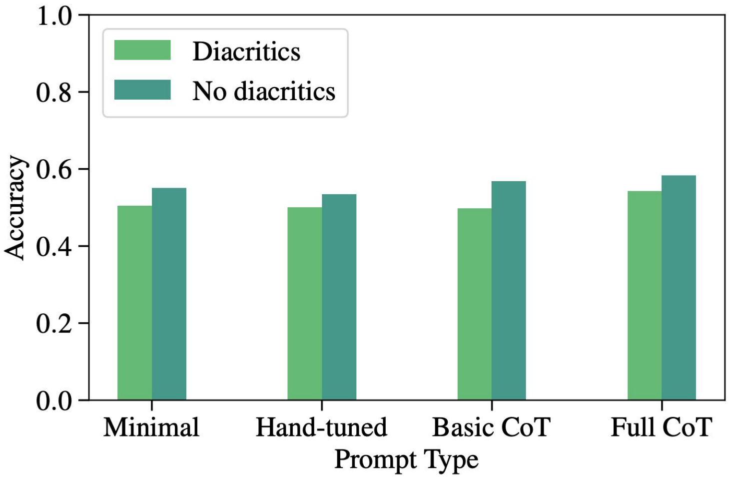
Orthography

We expose a frailty: LLMs do significantly better (4.8% higher absolute accuracy) when accent marks have been converted to ASCII!

Diacritics → ASCII

Ëjts yë' maxu'unk nexyejtpy.

`EUjts yeuq maxuqunk nexyejtpy.`



Conclusion

- LLMs show non-negligible abilities at few-shot multilingual reasoning.
- These abilities cannot purely be explained by data leakage.
- There is plenty of room for improvement: ModeLing can be used to measure progress in this area!

Acknowledgements

- Thanks to Lori Levin and Aleka Blackwell (NACLO) for helpful discussions.
- We would like to acknowledge our lead senior author, Dragomir Radev, who passed away during the preparation of this manuscript.