Prompting is not a substitute for probability measurements in large language models

Jennifer Hu
Harvard University

Roger Levy
MIT

Chapter 1

Modern language models refute Chomsky's approach to language

Steven T. Piantadosi[a,b]
[a]UC Berkeley, Psychology [b]Helen Wills Neuroscience Institute

Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi (2023).

Roni Katzir, Tel Aviv University

(What) Can Deep Learning Contribute to Theoretical Linguistics?

Author: Gabe Dupre Authors Info & Claims

Large Language Models Demonstrate the Potential of Statistical Learning in Language

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, Morten H. Christiansen

First published: 25 February 2023 | https://doi.org/10.1111/cogs.13256 | Citations: 2

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer D. Ullman
Department of Psychology
Harvard University
Cambridge, MA, 02138
tullman@fas.harvard.edu

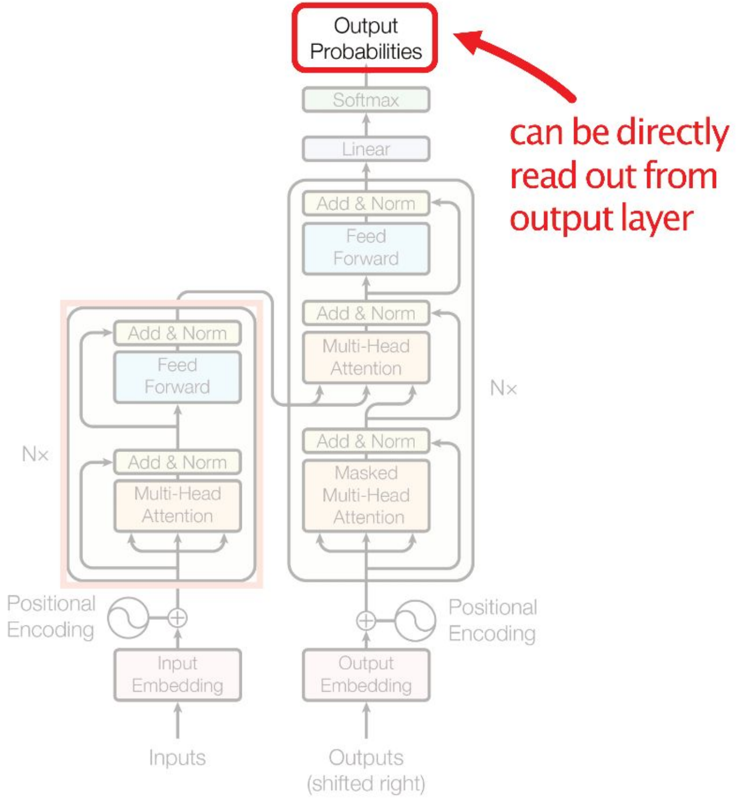
Slide credit: Jennifer Hu

How should we evaluate LLMs' linguistic abilities?

Slide credit: Jennifer Hu

The internal distribution

- Fundamental unit of LLM computation: P(token|context)



Vaswani et al. (2017)

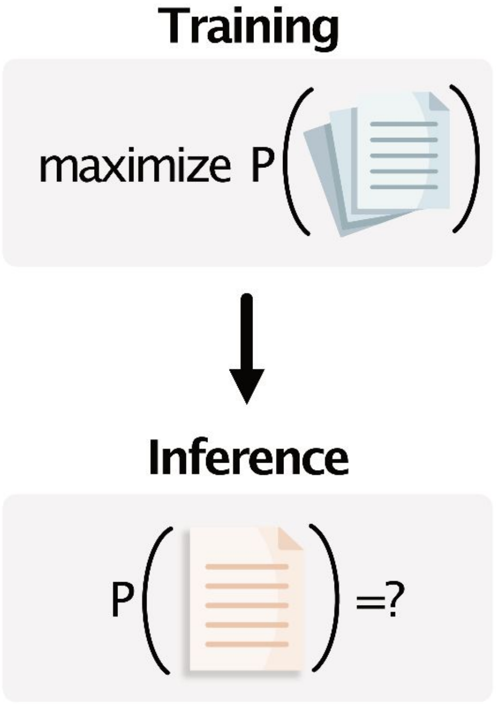
Slide credit: Jennifer Hu

The internal distribution

- Fundamental unit of LLM computation: P(token|context)
- This distribution reflects the model's **linguistic generalizations**:

  a generative model of the language seen during training…

  …which can be used to evaluate the likelihood of previously unseen strings



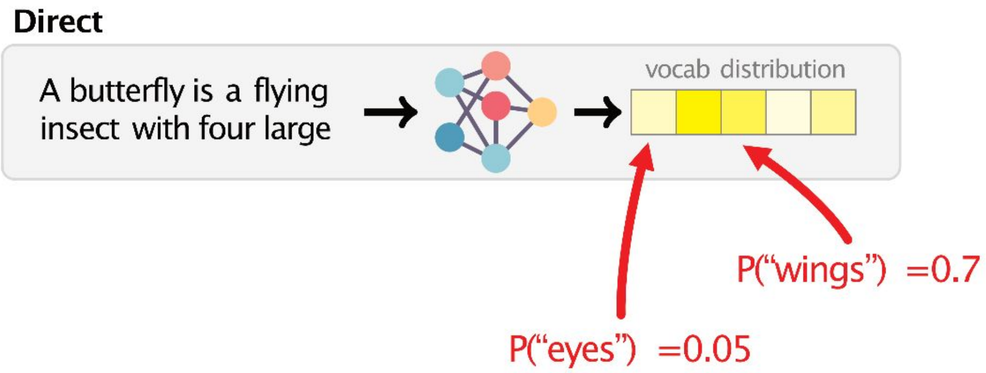
A new method: prompting

- Reveals new classes of emergent abilities in LLMs
  (Brown et al. 2020; Wei et al. 2022; Patel & Pavlick 2022; inter alia)
- Caveat: tests not only whether a model represents a certain generalization, but also whether the model can report the outcome of applying the generalization to the sentence in the prompt

Prompting tests a new emergent ability: **metalinguistic judgment**

Slide credit: Jennifer Hu

Example: next-wordprediction



Slide credit: Jennifer Hu

Example: next-wordprediction

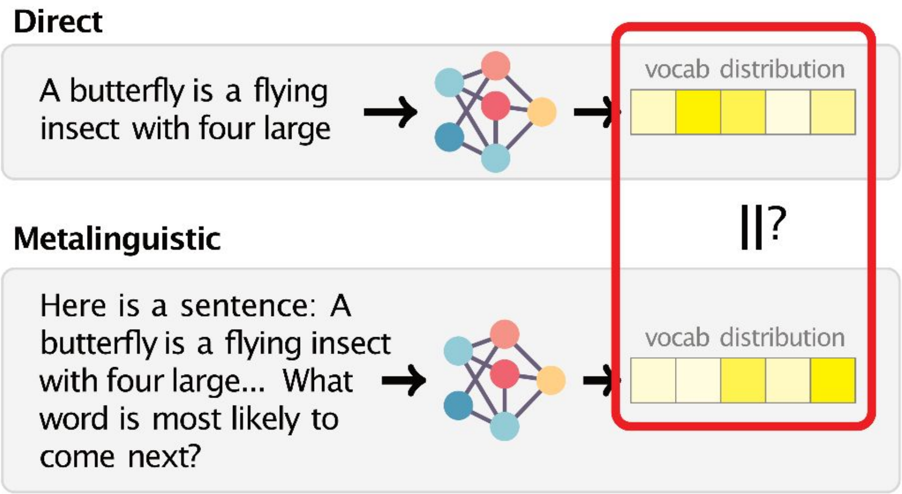
Direct

A butterfly is a flying insect with four large

Metalinguistic

Here is a sentence: A butterfly is a flying insect with four large... What word is most likely to come next?

Example: next-wordprediction



Slide credit: Jennifer Hu

Example: next-word prediction

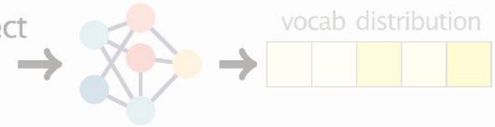
Direct

A butterfly is a flying insect with four large

||?

Metalinguistic

Here is a sentence: A butterfly is a flying insect with four large... What word is most likely to come next?



Example: sentence judgment

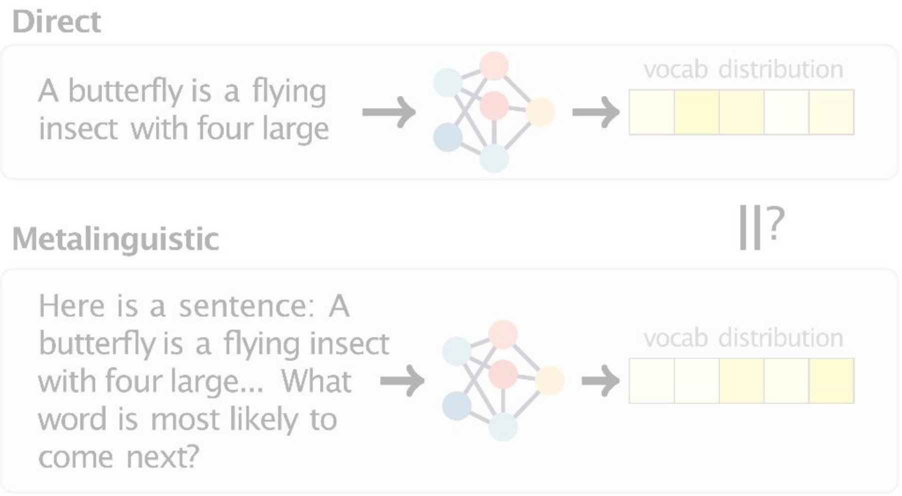
Direct

S_1 = Every child has studied. ✓

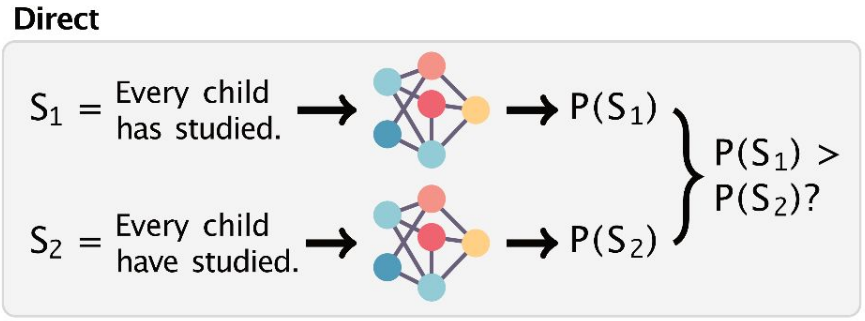
S_2 = Every child have studied. ✗

Slide credit: Jennifer Hu

Example: next-wordprediction



Example: sentence judgment



Slide credit: Jennifer Hu

Example: next-word prediction

Direct

A butterfly is a flying insect with four large → → vocab distribution

||?

Metalinguistic

Here is a sentence: A butterfly is a flying insect with four large... What word is most likely to come next? → → vocab distribution

Example: sentence judgment

Direct

S_1 = Every child has studied. → → $P(S_1)$

S_2 = Every child have studied. → → $P(S_2)$

$P(S_1) > P(S_2)$?

Metalinguistic

Here are two English sentences: 1) Every child has studied. 2) Every child have studied. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer. →  → vocab distribution

P("1") > P("2")?

Slide credit: Jennifer Hu

Contribution

In this paper, the authors evaluate the validity of metalinguistic prompting as a way of measuring LLMs' internal knowledge.

Two research questions:

1. How well do models perform under direct and metalinguistic evaluation methods?
2. How consistent are the metalinguistic methods with the direct method?

Four Experiment

Targeted ability	Task
Word prediction	Predict final word in a sentence
Semantic plausibility	Determine which word (of two options) is most likely, given preceding context
Syntax	Determine which sentence (of two options) is "better", in isolation
Syntax	Determine which sentence (of two options) is "better", given both options

Table 1: Overview of experiments in our study.

Direct Method: Computing probabilities of predicted tokens.

Zero-shot metalinguistic prompting: Ask a question or specify a task requiring a judgment about a linguistic expression.

LLMs

- Flan-T5 models:
  - small,
  - large,
  - XL
- GPT-3/3.5 models:
  - textcurie-001/GPT-3,
  - text-davinci-002/GPT-3.5,
  - textdavinci-003/GPT-3.5

Evaluation method

Accuracy evaluation:

- Compare the log probability of the predicted token.
- Pseudo log probability of the whole sentence.

Internal consistency between direct and metalinguistic evaluation:

Average correlation coefficient (Pearson’s r) between the item level **differentials** measured by the direct method and a particular metalinguistic prompting method.

Experiment 1: word prediction

- A simplified version of next word prediction.
- Predict the final word of a sentence.
- Datasets:
  - P18: 384 simple declarative sentences that state a fact about familiar concepts.
  - News: 222 sentence from recent news (title-first sentence).

Experiment 1(word prediction): prompt example

Type of prompt	Example
Direct	A butterfly is a flying insect with four large **wings**
MetaQuestionSimple	What word is most likely to come next in the following sentence? A butterfly is a flying insect with four large **wings**
MetaInstruct	You are a helpful writing assistant. Tell me what word is most likely to come next in the following sentence: A butterfly is a flying insect with four large **wings**
MetaQuestionComplex	Here is the beginning of an English sentence: A butterfly is a flying insect with four large... What is the best next word? Answer: **wings**

Table 2: Example prompts for Experiment 1. Region where we measure probability is marked in **boldface**. Ground-truth sentence continuations are shown in blue.

Experiment 2: semantic plausibility

- Judge which of two words is a more likely continuation of a sentence.
- Assess knowledge of semantic plausibility.
- Dataset:
  - Minimal pair: 395 minimal sentence pair. Each pair consist of two sentences that differ only in the final words.
  - Example: The archer released the **arrow/interview**.

Experiment 2(semantic plausibility): prompt example

Type of prompt	Example
Direct	The archer released the {**arrow**, **interview**}
MetaQuestionSimple	What word is most likely to come next in the following sentence (arrow, or interview)? The archer released the {**arrow**, **interview**}
MetaInstruct	You are a helpful writing assistant. Tell me what word is most likely to come next in the following sentence (arrow, or interview?): The archer released the {**arrow**, **interview**}
MetaQuestionComplex	Here is the beginning of an English sentence: The archer released the... What word is more likely to come next: arrow, or interview? Answer: {**arrow**, **interview**}

Table 3: Example prompts for Experiment 2. Region where we measure probability is marked in **boldface**. Semantically plausible continuations are shown in blue; implausible in red.

Experiment 3a: Sentence judgment (isolated)

- Evaluate models' ability to judge whether a sentence is a "good" sentence of English.
- A good and bad sentence is evaluated separately.
- Dataset:
  - Minimal pair dataset of english grammatical syntax.
    - SyntaxGym
    - BLiMP

Experiment 3a(Sentence judgment) prompt example

Type of prompt	Example
Direct	{**Every child has studied**, **Every child have studied**}
MetaQuestionSimple	Is the following sentence a good sentence of English? Every child has studied. Respond with either Yes or No as your answer. {Yes, No}
MetaInstruct	You are a helpful writing assistant. Tell me if the following sentence is a good sentence of English. Every child has studied. Respond with either Yes or No as your answer. {Yes, No}
MetaQuestionComplex	Here is a sentence: Every child has studied. Is the sentence a good sentence of English? Respond with either Yes or No as your answer. Answer: {Yes, No}

(a)

Experiment 3b: Sentence comparison

- Measure models' syntactic judgments.
- However, instead of presenting the model with sentences in isolation, the experiment present the model with both sentence of a minimal pair.
- Dataset:
  - Same as experiment 3a (SyntaxGym, BLiMP)

Experiment 3b(Sentence comparison) prompt example

Type of prompt	Example
Direct	{**Every child has studied**, **Every child have studied**}
MetaQuestionSimple	Which sentence is a better English sentence? 1) Every child has studied. 2) Every child have studied. Respond with either 1 or 2 as your answer. {1, 2}
MetaInstruct	You are a helpful writing assistant. Tell me which sentence is a better English sentence. 1) Every child has studied. 2) Every child have studied. Respond with either 1 or 2 as your answer. {1, 2}
MetaQuestionComplex	Here are two English sentences: 1) Every child have studied. 2) Every child has studied. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer. Answer: {1, 2}

(b)

Task Performance

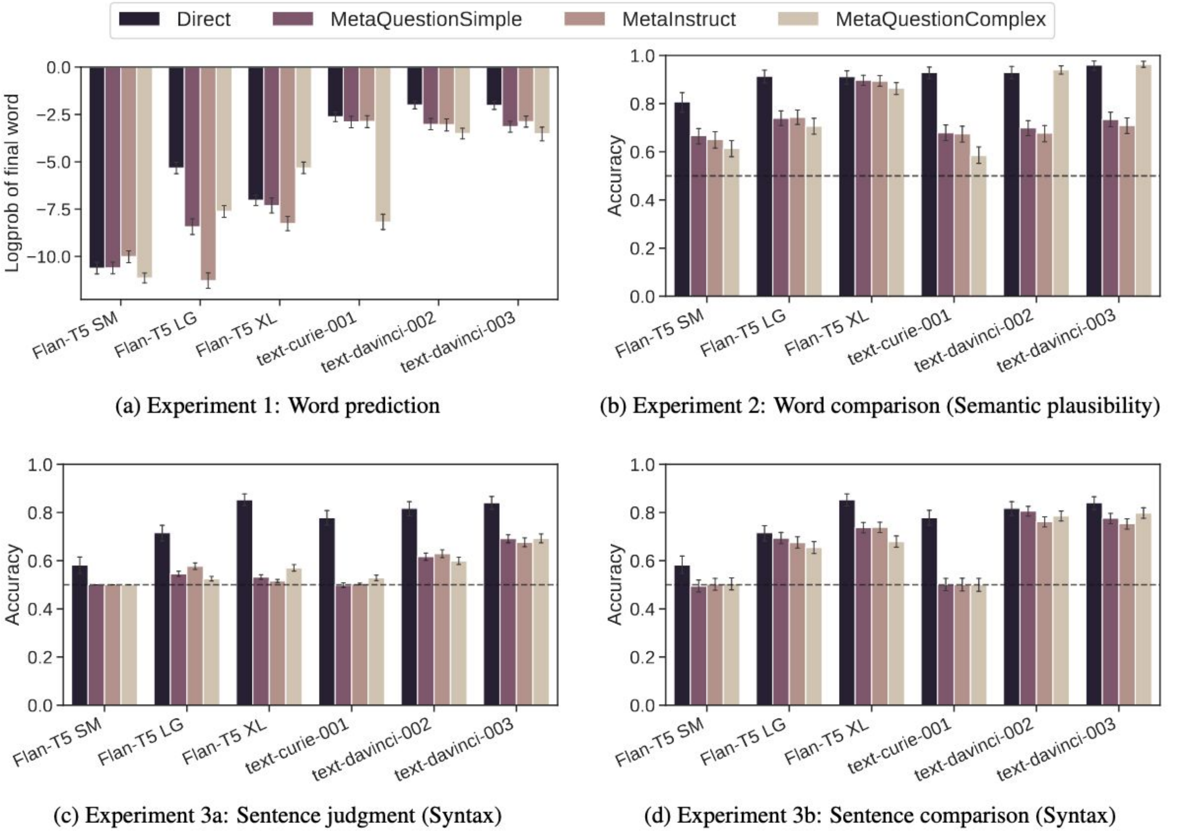


Figure 2: **Task performance: Direct probability measurements generally outperform metalinguistic prompts.** (a) Log probability assigned to ground-truth sentence continuation, averaged over items and datasets. (b) Proportion of items where model prefers semantically plausible continuation over implausible continuation. (c)-(d) Proportion of items where model prefers grammatical sentence over ungrammatical sentence in minimal pair, averaged over datasets. Error bars denote bootstrapped 95% CIs. Dashed lines indicate random baseline.

Task Performance

Metalinguistic judgments are not the same as direct measurements.

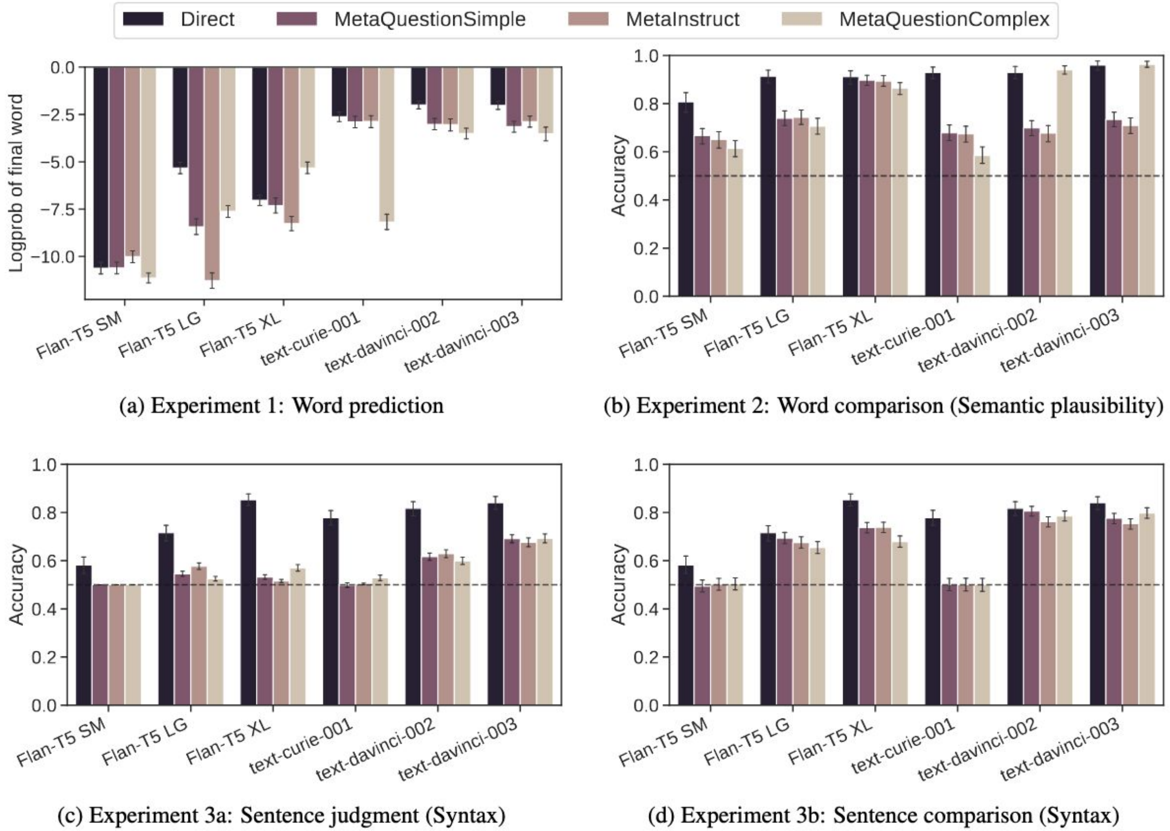
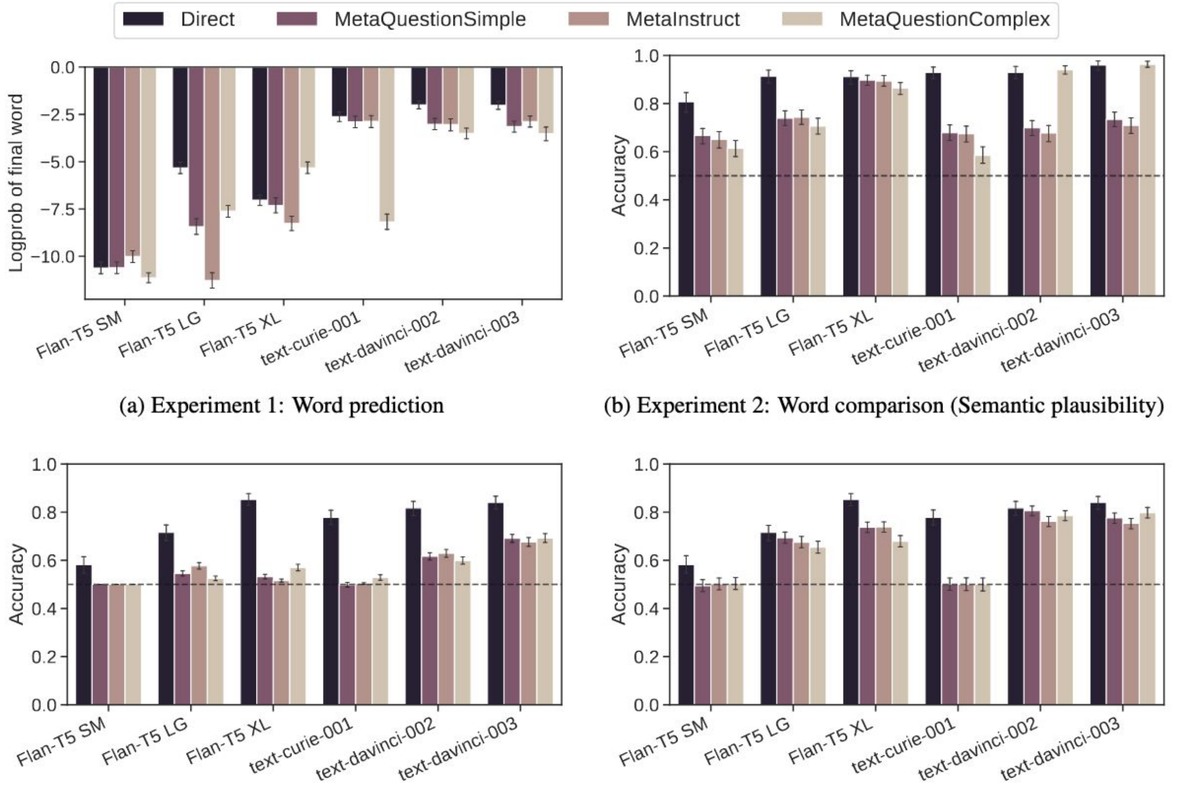


Figure 2: **Task performance: Direct probability measurements generally outperform metalinguistic prompts.** (a) Log probability assigned to ground-truth sentence continuation, averaged over items and datasets. (b) Proportion of items where model prefers semantically plausible continuation over implausible continuation. (c)-(d) Proportion of items where model prefers grammatical sentence over ungrammatical sentence in minimal pair, averaged over datasets. Error bars denote bootstrapped 95% CIs. Dashed lines indicate random baseline.

Task Performance

Direct measurements generally perform ≥ metalinguistic methods.



(a) Experiment 1: Word prediction

(b) Experiment 2: Word comparison (Semantic plausibility)

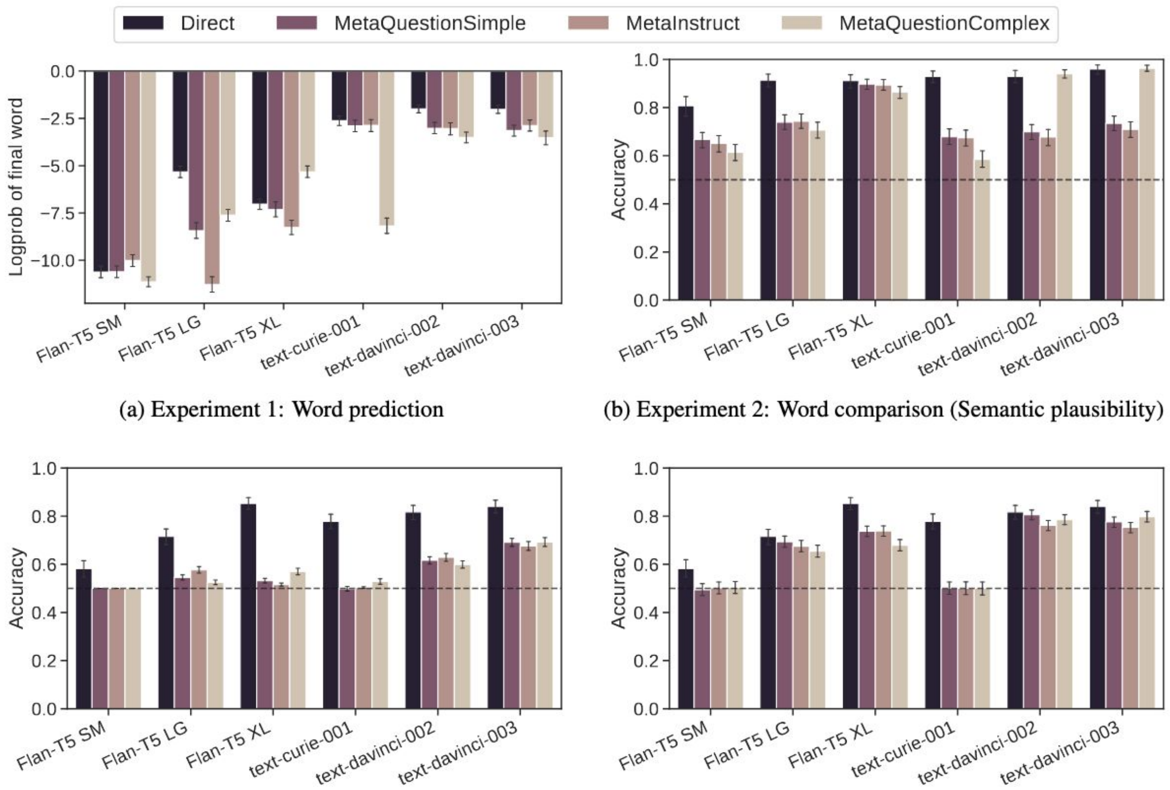
(c) Experiment 3a: Sentence judgment (Syntax)

(d) Experiment 3b: Sentence comparison (Syntax)

Figure 2: **Task performance: Direct probability measurements generally outperform metalinguistic prompts.** (a) Log probability assigned to ground-truth sentence continuation, averaged over items and datasets. (b) Proportion of items where model prefers semantically plausible continuation over implausible continuation. (c)-(d) Proportion of items where model prefers grammatical sentence over ungrammatical sentence in minimal pair, averaged over datasets. Error bars denote bootstrapped 95% CIs. Dashed lines indicate random baseline.

Task Performance

Minimal pairs help reveal models' generalization capacities.



(a) Experiment 1: Word prediction

(b) Experiment 2: Word comparison (Semantic plausibility)

(c) Experiment 3a: Sentence judgment (Syntax)

(d) Experiment 3b: Sentence comparison (Syntax)

Figure 2: **Task performance: Direct probability measurements generally outperform metalinguistic prompts.** (a) Log probability assigned to ground-truth sentence continuation, averaged over items and datasets. (b) Proportion of items where model prefers semantically plausible continuation over implausible continuation. (c)-(d) Proportion of items where model prefers grammatical sentence over ungrammatical sentence in minimal pair, averaged over datasets. Error bars denote bootstrapped 95% CIs. Dashed lines indicate random baseline.

Internal consistency

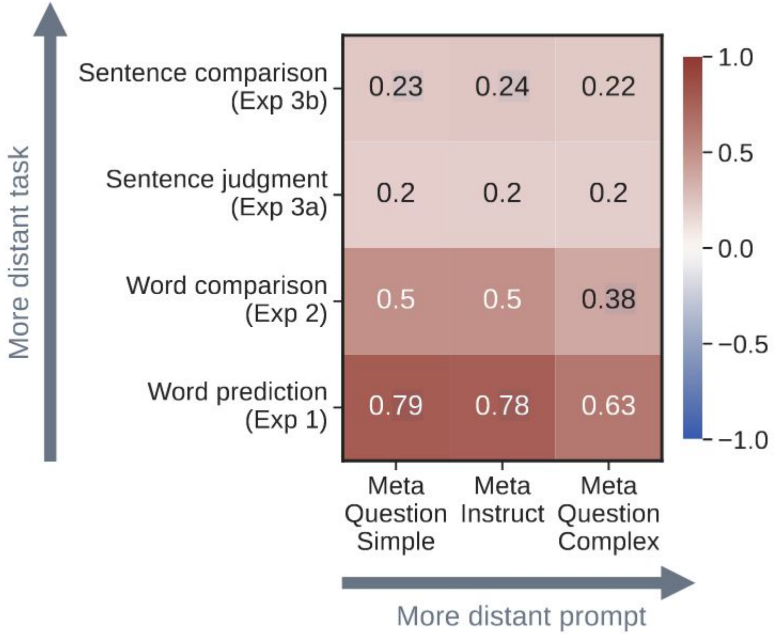


Figure 3: **Internal consistency: Correlation between metalinguistic and direct responses gets weaker as prompts become less direct.** Pearson r correlation between response magnitudes (averaged over models and datasets) measured by direct prompts versus each metalinguistic prompt. See Appendix C for more details.

Internal consistency

Consistency gets worse as we get further from direct measurement of next-word probabilities.

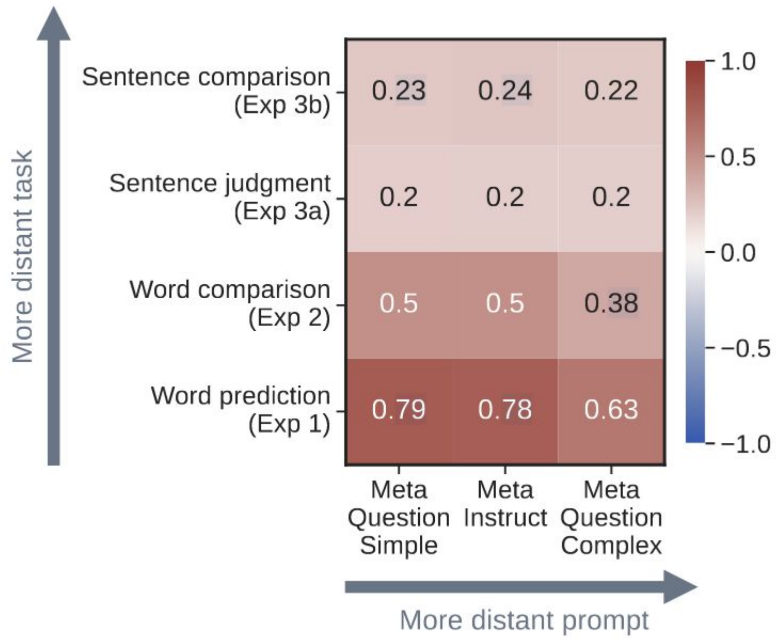


Figure 3: **Internal consistency: Correlation between metalinguistic and direct responses gets weaker as prompts become less direct.** Pearson r correlation between response magnitudes (averaged over models and datasets) measured by direct prompts versus each metalinguistic prompt. See Appendix C for more details.

Discussion

- Taken together, their findings suggest that negative results relying on metalinguistic prompts cannot be taken as conclusive evidence that an LLM lacks a particular linguistic generalization.
- These findings suggest a possible basis for a **competence** performance distinction in LLMs: namely, the distinction between the information encoded in a model's isolated-sentence string probability distribution versus the model's **behavioral** responses to prompts.
- Their results also highlight the value that is lost as researchers move toward interacting with LLMs through closed APIs, where access to models' underlying probability distributions is limited.