Carnegie Mellon University
Language
Technologies
Institute

Verbing Wugs

Evaluating Morphological Generalization in Large Language Models

David R. Mortensen

March 7, 2024

ETH Zürich

Large Language Models have been suggested to have human-like linguistic behavior

Large Language Models have been suggested to have human-like linguistic behavior

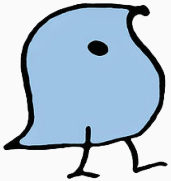
Is this true of their morphological behavior?

Large Language Models have been suggested to have human-like linguistic behavior

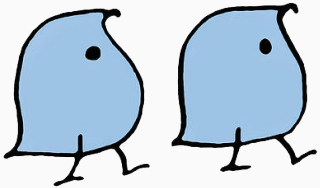
Is this true of their morphological behavior? And can we study this question in the same way psycholinguists have studied morphological performance in humans?

How We Know Humans are Capable of Morphological Generalization

In 1958, Jean Berko-Gleason reported what was possibly the most impactful experiment in the history of psycholinguistics.



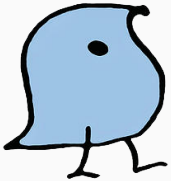
THIS IS A WUG'



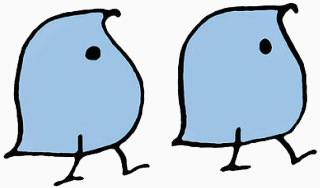
NOW THERE IS ANOTHER ONE
THERE ARE TWO OF THEM
THERE ARE TWO _____

How We Know Humans are Capable of Morphological Generalization

In 1958, Jean Berko-Gleason reported what was possibly the most impactful experiment in the history of psychoLinguistics.



THIS IS A WUG'

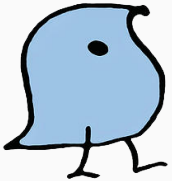


NOW THERE IS ANOTHER ONE
THERE ARE TWO OF THEM
THERE ARE TWO _____

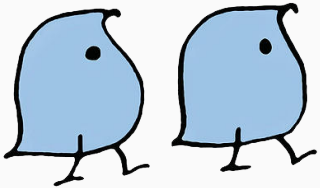
This paper is about doing psycholinguistics on large language models.

How We Know Humans are Capable of Morphological Generalization

In 1958, Jean Berko-Gleason reported what was possibly the most impactful experiment in the history of psycholinguistics.



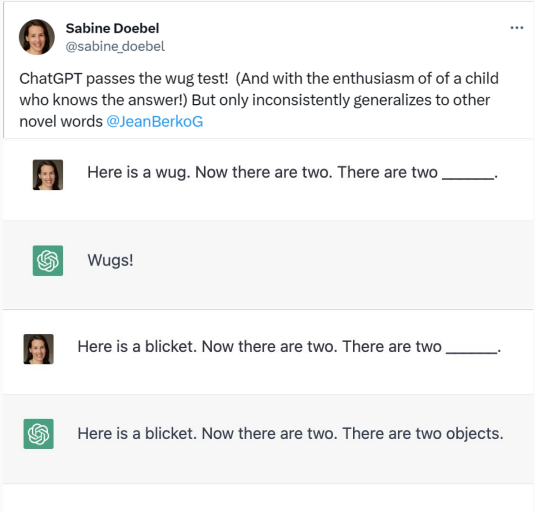
THIS IS A WUG'



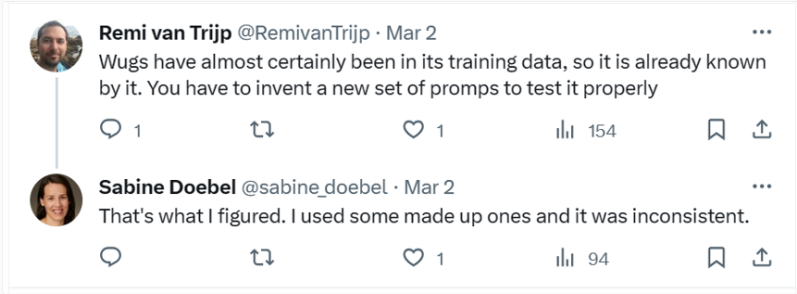
NOW THERE IS ANOTHER ONE
THERE ARE TWO OF THEM
THERE ARE TWO _____

This paper is about doing psycholinguistics on large language models. The focus is on tests of implicit (rather than explicit) competence.

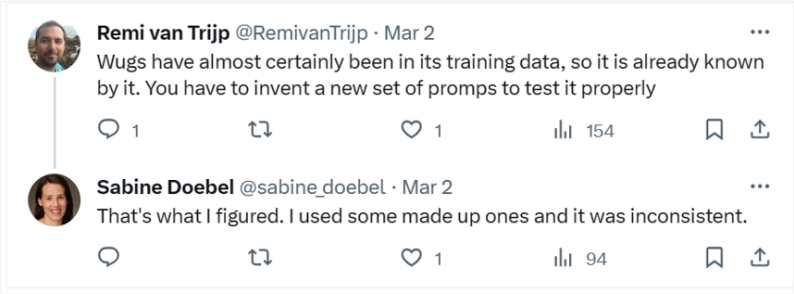
Can ChatGPT Pass the Wug Test?



Can ChatGPT Pass the Wug Test?



Can ChatGPT Pass the Wug Test?



Without having seen Remi van Trijp's tweet, we did what he suggested [EMNLP'23].

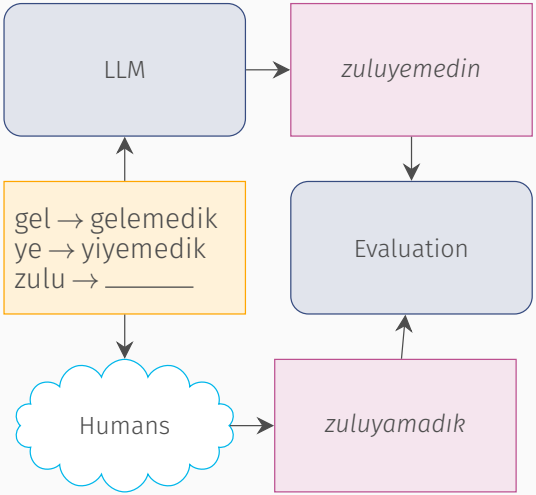
Counting the Bugs in ChatGPTs Wugs [EMNLP'23]

The First Systematic Wug-Test of ChatGPT

- Completely new nonce words (“wug words”) $\rightarrow$ no contaminated training data
- Four languages: English, German, Tamil, Turkish
  - High- and low-resource
  - Typologically diverse morphological systems
  - Different phenomena (past tense of verbs, plurals of nouns, ...)
- Judgments collected from human annotators, forms ranked by frequency of production

English: veed $\rightarrow 18 \times$ veeded, $6 \times$ ved, $3 \times$ veed, $1 \times$ vode
German: Sater $\rightarrow 7 \times$ Sater, $4 \times$ Saters, $4 \times$ Satere

Evaluation Setup



- Zero-shot, one-shot or few-shot (i.e., one per inflection class) scenarios for ChatGPT
- Evaluation with acc@1, @3 and @5 against the ranked list of human judgments
- Two prompt setups
  - LONG: explanation and sentence with blank
  - SHORT: only explanation

Dataset

Lang.	Train	Dev	Test	Wug test	Phenomenon
English	10,000	1,000	1,000	50	past tense
German	10,000	1,000	1,000	174	plural number
Tamil	1,541	368	—	123	past tense
Turkish	8,579	851	846	40	accusative singular

Annotator Accuracy

	Accuracy (%)		
Lang.	@1	@3	@5
English	67.14 ± 17.76	85.29 ± 13.06	87.64 ± 12.13
German	63.05 ± 12.62	83.80 ± 10.57	87.88 ± 10.34
Tamil	37.09 ± 26.39	43.85 ± 26.95	43.85 ± 26.95

Baselines

ARL Affix Rule Learner. from Simple non-neural system (Liu and Mao, 2016) use edit distance to discover affix rules.

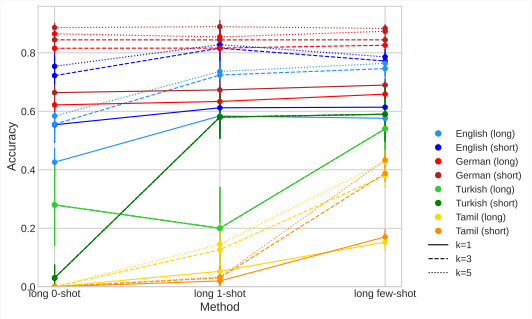
MinGen Minimal Generalization Learner. System based on a simplified form of Albright and Hayes (2002) from (Wilson and Li, 2021)

FIT Feature Invariant Transformer. Character-level transformer Wu et al. (2021) for feature-guided transduction.

PPI Principle Parts for Inflection. Morphological inflection as "paradigm cell filling" (Liu and Hulden, 2020).

AED Analogical Encoder-Decoder. Encoder-decoder architecture augmented with pre-compiled analogical patterns for inflecting nonce words Calderone et al. (2021), following up on Albright and Hayes (2003) and Kirov and Cotterell (2018)..

Overall Performance: GPT-3.5



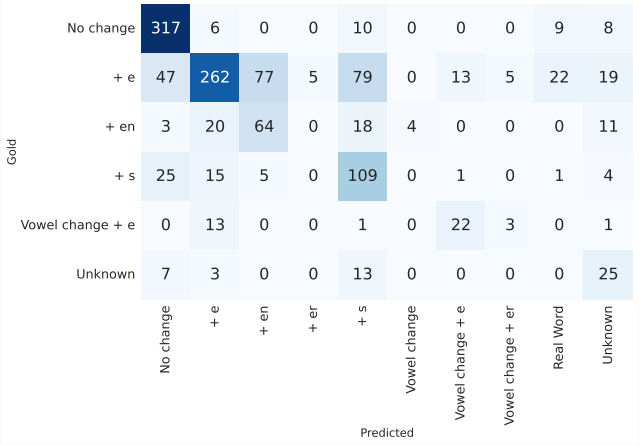
- English and German perform better than Tamil and Turkish
- More shots → better performance for Tamil and Turkish
- Even for English, performance is worse than 3 out of 5 baseline morphological inflectors

Overall Performance

Method	English	German	Tamil	Turkish
ARL	**100.00**	**94.25**	61.48	60.00
MinGen	62.00	64.37	49.18	40.00
FIT	98.00 ± 1.26	92.87 ± 0.74	**63.28 ± 3.36**	67.00 ± 4.58
PPI	94.60 ± 2.54	85.98 ± 5.91	55.33 ± 1.84	**68.00 ± 4.00**
AED	57.60 ± 6.62	48.51 ± 5.45	58.69 ± 5.46	56.00 ± 4.90
long 0-shot	58.40 ± 5.28	86.49 ± 1.07	0.00	28.00 ± 14.00
long 1-shot	73.60 ± 6.97	85.42 ± 2.52	14.52 ± 7.48	20.00 ± 14.14
long few-shot	76.40 ± 4.45	87.36 ± 2.37	42.70 ± 3.96	54.00 ± 10.20
short 0-shot	75.40 ± 5.87	88.62 ± 1.64	0.00	3.00 ± 4.58
short 1-shot	82.80 ± 5.60	88.94 ± 2.35	3.28 ± 3.99	58.00 ± 7.48
short few-shot	78.60 ± 2.84	88.33 ± 1.15	43.36 ± 3.12	59.00 ± 9.43

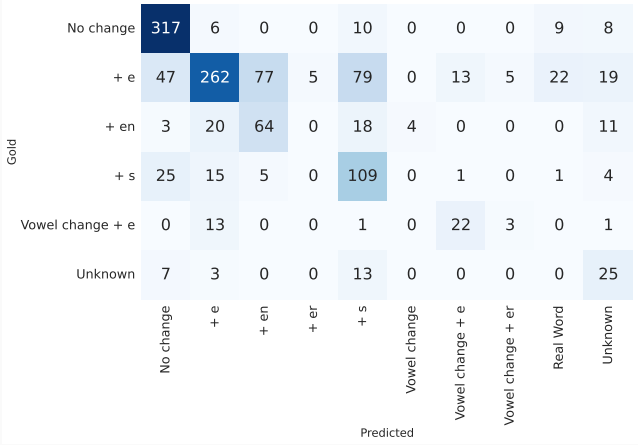
ARL, the simplest method, always beats GPT-3.5.

Overgeneralization in German



Main source of error in German is overgeneralization to the most productive plural morphemes, *+en* and *+s*

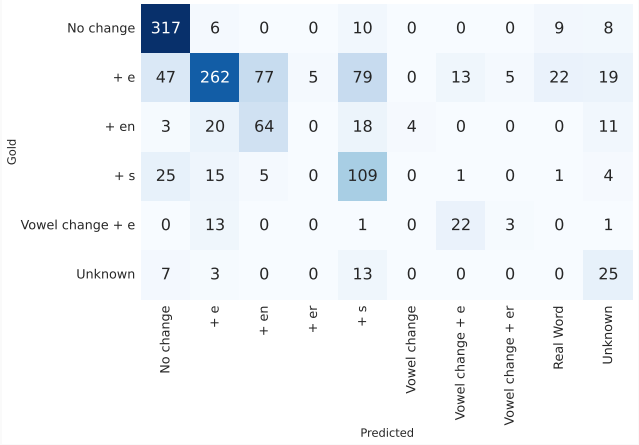
Overgeneralization in German



Main source of error in German is overgeneralization to the most productive plural morphemes, *+en* and *+s*

Amplification of biases in the training data!

Overgeneralization in German

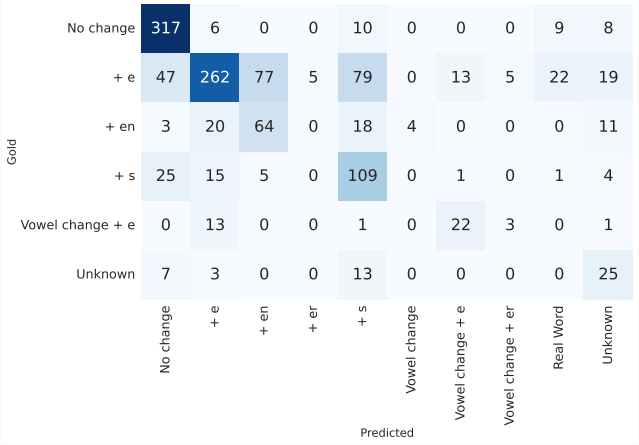


Main source of error in German is overgeneralization to the most productive plural morphemes, *+en* and *+s*

Amplification of biases in the training data!

But the situation is different in English, were “overgeneralizing” would always yield the response most annotators provided.

Overgeneralization in German



Main source of error in German is overgeneralization to the most productive plural morphemes, *+en* and *+s*

Amplification of biases in the training data!

But the situation is different in English, were “overgeneralizing” would always yield the response most annotators provided.

Why this difference?

Real Word Bias

Many errors are not human-like

veed	→	**viewed**, **videoed**, **veered**, veedoed
trin	→	**trained**
bebit	→	**drained**, **drank**, bebitted, bebit, bebit, **bet**, bebitted, bebit, bdrank

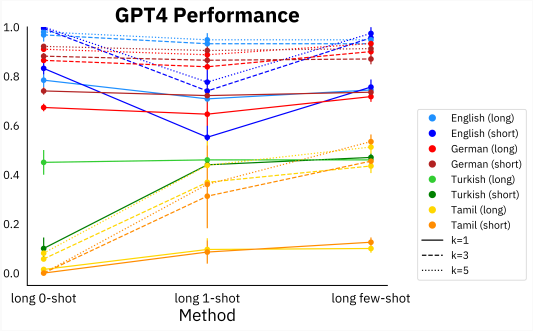
Real Word Bias

Many errors are not human-like

veed	→	**viewed**, **videoed**, **veered**, veedoed
trin	→	**trained**
bebit	→	**drained**, **drank**, bebitted, bebit, bebit, **bet**, bebitted, bebit, bdrank

These often involve the models gravitating towards orthographically-similar real words (in English, specifically).

What about GPT-4?



English long prompt and Tamil short prompt as examples

- English near-perfect for 0-shot and few-shot, but worse than GPT-3.5 for 1-shot → highly suggestible, generalises from the one shot
- Tamil slightly better for 0-shot

The Upshot of this Experiment:

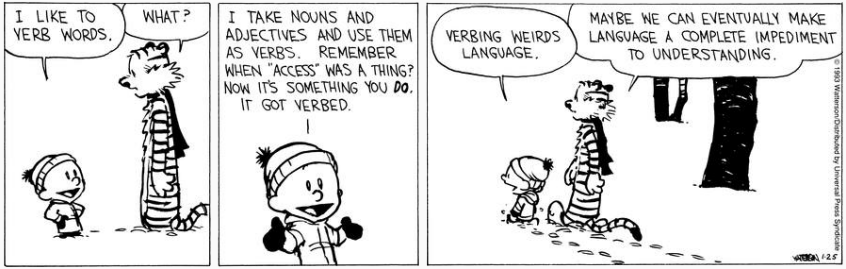
The OpenAI models are getting better at morphological generalization, but they are still not better than the simplest (supervised) baseline

Verbing Weirds Language (Models) [LREC-COLING'24]

But there are Other Kinds of Morphological Generalization

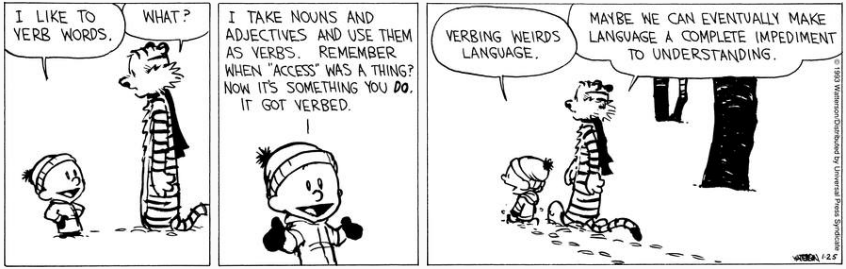
Zero-Derivation

Zero-Derivation



Are LLMs like ChatGPT, Llama 2, Mistral, and Falcon robust to conversion of words to non-prototypical parts of speech (frequently observed when people are Englishing)?

Zero-Derivation



Are LLMs like ChatGPT, Llama 2, Mistral, and Falcon robust to conversion of words to non-prototypical parts of speech (frequently observed when people are Englishing)?

Motivation

Conversion or zero-derivation is pervasive in English (and many other languages): a word with one (or more) prototypical parts of speech is used in a context that calls for **another** part of speech.

You can "verb" various parts of speech in English:

Adjective His hair has begun to *gray*.

Mass Noun If you don't want to *water* the plants, please *coffee* the graduate students instead.

Count Noun The fascist tried to *knife* me in the back.

Note: *to gray*, *to knife* are established usage; *to coffee* is not.

Data

transitive verbs 42 words listed in UniMorp's English language set as verbs but not nouns.

intransitive verbs 42 words listed in UniMorph as verbs but not nouns

mass nouns 51 words listed in UniMorph as nouns but not verbs

count nouns 79 words listed in UniMorph as nouns but not verbs

nonce words 49 nonce words generated automatically by Unipseudo based on the list of 59 most frequent mono-morphemic nouns and verbs of length 6 listed in UniMorph either as noun but not verb or vice versa, manually culled to remove words that were to similar to or too distant from existing English words.

All lexical sets were manually curated by a native-speaker linguist.

A Natural Language Inference Task for Zero Derivation

We tested the ability of LLMs to generalize about zero-derivation by forcing them to answer questions that required construing the same orthographic word as having non-prototypical (and prototypical) parts of speech:

Prototypical: If I **thrive** daily, do I **thrive** every day?
Non-prototypical: If I **health** daily, do I **health** every day?
Nonce: I **volice** daily, do I **volice** every day?

We were looking for differences between the non-prototypical condition, on the one hand, and the prototypical and nonce conditions in how their responses matched our reference responses

A Natural Language Inference Task for Zero Derivation

We tested the ability of LLMs to generalize about zero-derivation by forcing them to answer questions that required construing the same orthographic word as having non-prototypical (and prototypical) parts of speech:

Prototypical: If I **thrive** daily, do I **thrive** every day?
Non-prototypical: If I **health** daily, do I **health** every day?
Nonce: I **volice** daily, do I **volice** every day?

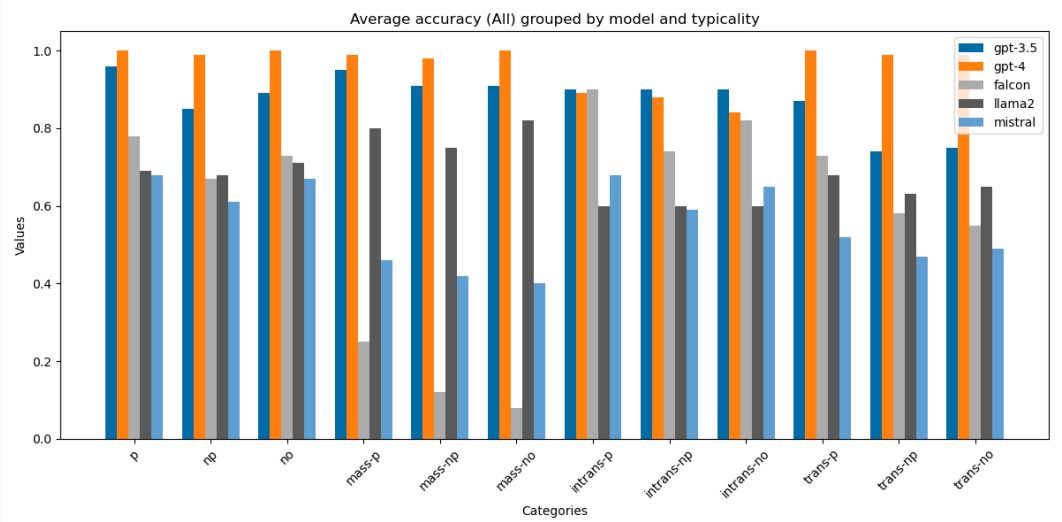
We were looking for differences between the non-prototypical condition, on the one hand, and the prototypical and nonce conditions in how their responses matched our reference responses (In the examples, “Yes”).

Hypotheses

1. LLMs will answer less consistently with the reference in the non-prototypical condition than the prototypical condition
2. LLMs will answer less consistently with the reference in the nonce condition that the non-prototypical condition
3. There will be a correlation between performance on the prototypical conditions and the other two conditions
4. The difference in LLM performance can be explained primarily by the size of the models.

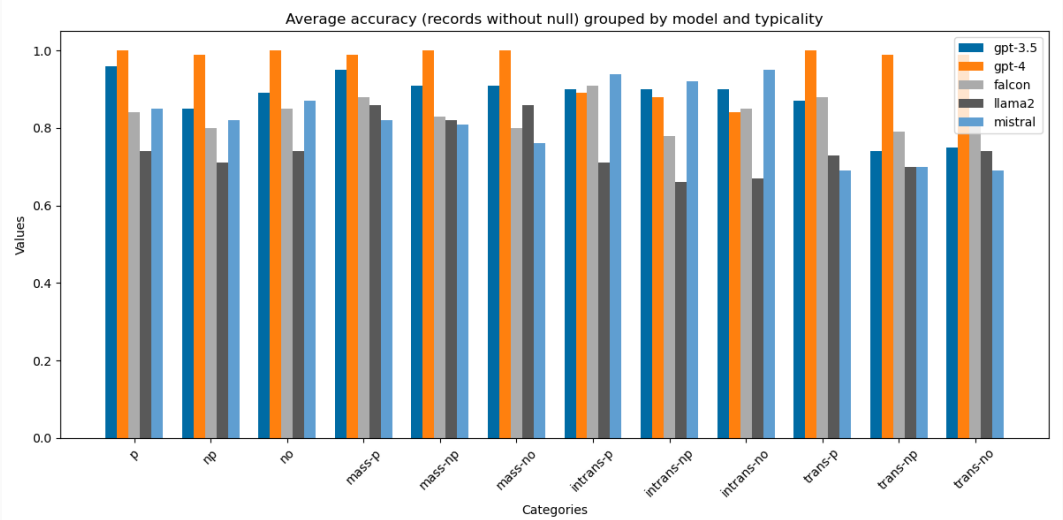
Results: with Nulls

GPT-4 > GPT-3.4 > Falcon-40B > Llama-70B, Mistral-7B



Results: without Nulls

Part, but not all, of the difference can be explained by instruction-following ability



Regression Analysis

- **Model:** Logistic Regression
- **Factors:**
  - Prototypical part of speech
  - Model type
  - Prototypicality of filler given frame
  - Answer (yes, no, “null”)
- **Results:**
  - All factors significant ($p < 0.01$)
  - Answer type as strongest predictor.
  - Prototypical part of speech is also a strong predictor

Model Type is also a Significant Predictor

- GPT-4 is the best, followed by GPT-3.5
- The best open model is Mistral 7B, even though it is smaller than Llama 2 70B and Falcon 40B.
- What drags Falcon down seems to be its reluctance to follow instructions (not generalization ability per se)
- **Performance on these tasks is not a function of model size, but of other aspects of their training.**

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance — Supported

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance — Supported

non-prototypical performance > nonce performance

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance	Supported
non-prototypical performance > nonce performance	Not supported

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance	Supported
non-prototypical performance > nonce performance	Not supported
Correlation between prototypical, non-prototypical, nonce performance	

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance	Supported
non-prototypical performance > nonce performance	Not supported
Correlation between prototypical, non-prototypical, nonce performance	Supported

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance	Supported
non-prototypical performance > nonce performance	Not supported
Correlation between prototypical, non-prototypical, nonce performance	Supported
Difference between model size accounts for difference in performance	

How did the Hypotheses Hold Up?

prototypical performance > non-prototypical performance	Supported
non-prototypical performance > nonce performance	Not supported
Correlation between prototypical, non-prototypical, nonce performance	Supported
Difference between model size accounts for difference in performance	Not supported

Conclusioning

- GPT-3.5 and (especially GPT-4) are very good at the verbing task, in part—by not completely—because they follow instructions well
- The open models lag behind, but not in a way that can be explained by model size (Mistral-7B is roughly as good as Llama-70B) and Falcon-40B is better than either.
- Unlike inflection, existing language models are able to perform this task well.

Looking Forward

Current LLMs do not display exactly the degree of morphological generalization that humans do.

Looking Forward

Current LLMs do not display exactly the degree of morphological generalization that humans do.

However, this appears to be a limitation of degree, and not in kind.

Looking Forward

Current LLMs do not display exactly the degree of morphological generalization that humans do.

However, this appears to be a limitation of degree, and not in kind.

GPT-4 has near perfect performance on the verbing task, and approaches the human ceiling on the wug task even though GPT-3 did not.

Looking Forward

Current LLMs do not display exactly the degree of morphological generalization that humans do.

However, this appears to be a limitation of degree, and not in kind.

GPT-4 has near perfect performance on the verbing task, and approaches the human ceiling on the wug task even though GPT-3 did not. However, the ultimate test of such models is whether they can achieve human-like performance with human-like levels of training data.

Thanks to my Many Collaborators

Leonie Weissweiler

Valentin Hofmann

Anjali Kantharuban

Anna Cai

Valentina Izrailevitch

Haofei Yu

Abhishek Vijayakumar

Atharva Kulkarni

Lorenzo Xiao

Anubha Kabra

Ritam Dutt

Amey Hengle

Kemal Oflazer

Questions?

References

Adam Albright and Bruce Hayes. 2002. Modeling english past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, page 58–69, USA. Association for Computational Linguistics.

Adam Albright and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2-3):150–177.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 274–282, Online. Association for Computational Linguistics.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 153–161, Online. Association for Computational Linguistics.

Ling Liu and Lingshuang Jack Mao. 2016. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 36–40, Berlin, Germany. Association for Computational Linguistics.

Colin Wilson and Jane S.Y. Li. 2021. Were we there already? Applying minimal generalization to the SIGMORPHON-UniMorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 283–291, Online. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.