Carnegie Mellon University
Language
Technologies
Institute

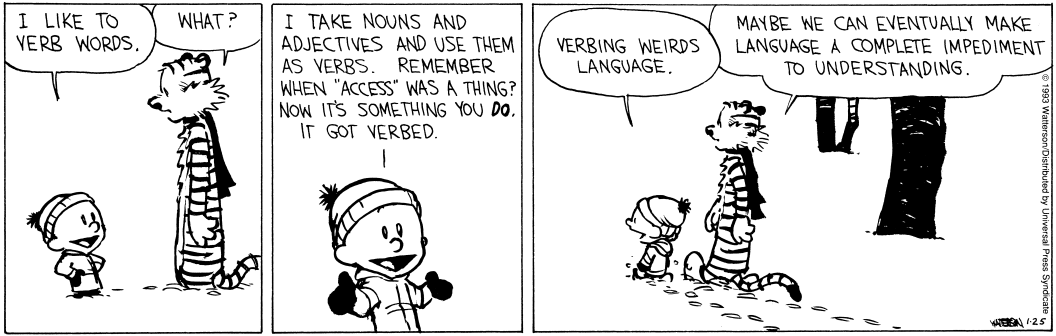
Verbing Weirds Language (Models)

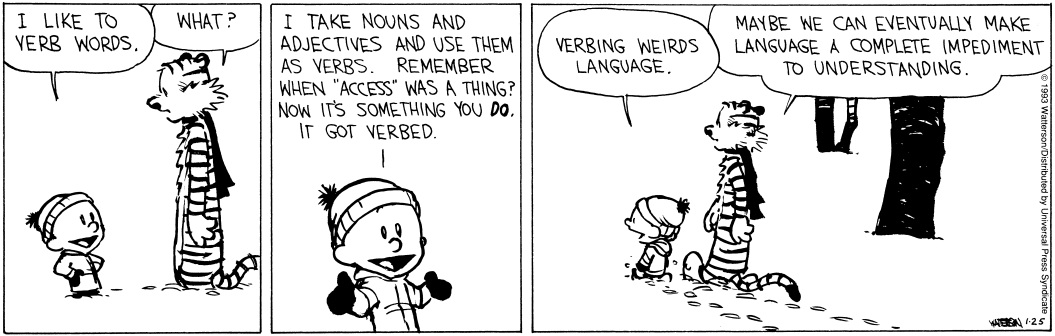
Evaluation of English Zero-Derivation in Five LLMs

David R. Mortensen, Valentina Izrailevitch, Yunze Xiao, Hinrich Schütze, and Leonie Weissweiler

May 2, 2024

Carnegie Mellon University and LMU Munich


I LIKE TO VERB WORDS.
WHAT?
I TAKE NOUNS AND ADJECTIVES AND USE THEM AS VERBS. REMEMBER WHEN "ACCESS" WAS A THING? NOW IT'S SOMETHING YOU DO. IT GOT VERBED.
VERBING WEIRDS LANGUAGE.
MAYBE WE CAN EVENTUALLY MAKE LANGUAGE A COMPLETE IMPEDIMENT TO UNDERSTANDING.
WATTERSON 12-2
© 1993 Watterson/Distributed by Universal Press Syndicate



English speakers like to verb words!

$$\text{Verbing} \subset (\text{Conversion} = \text{Zero-Derivation})$$

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.

Research Question

Are LLMs like **GPT-3.5**, **GPT-4**, **Llama2**, **Mistral**, and **Falcon** robust to conversion of words to non-prototypical parts of speech (as is seen when people English)?

Data

We curated five word lists:

Category	Num	UniMorph Noun	UniMorph Verb
transitive verbs	42	✗	✓
intransitive verbs	42	✗	✓
mass nouns	51	✓	✗
count nouns	79	✓	✗
nounce words	49	✗	✗

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: If 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, on the other, 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:

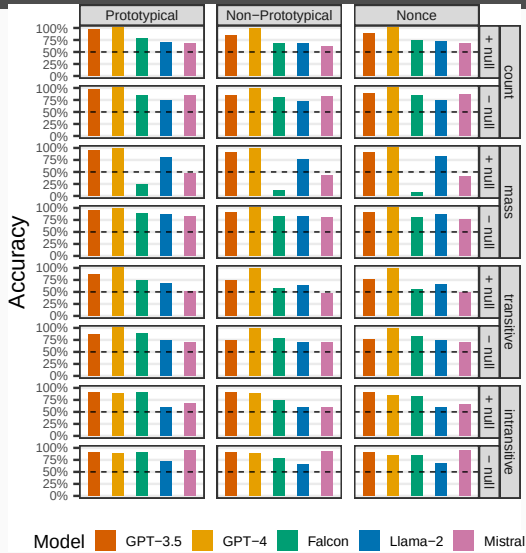
Prototypical: If I **thrive** daily, do I **thrive** every day?
Non-prototypical: If I **health** daily, do I **health** every day?
Nonce: If I **volice** daily, do I **volice** every day?

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

Hypotheses

1. LLMs answer less consistently with the reference in the non-prototypical condition than the prototypical condition
2. LLMs answer less consistently with the reference in the nonce condition than the non-prototypical condition
3. There is 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



- GPT-4 nearly perfect across all frame types except INTRANSITIVE
- GPT-3.5 performs similarly
- Falcon performs very well with INTRANSITIVE frames but poorly with MASS NOUN frames
- Llama performs very well on MASS frames and Mistral performs very well on INTRANSITIVE frames but they are otherwise comparable
- Prototypical > Non-Prototypical, Nonce

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 Falcon, even though it is smaller than Llama2 70B.
- 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—but 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 if you factor out null responses).
- 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—in the case of conversion—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—in the case of conversion—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-3.5 does not.

Looking Forward

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

However—in the case of conversion—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-3.5 does not. However, the ultimate test of such models is whether they can achieve human-like performance with human-like levels of training data, a subject of future work.

Thanks to my Collaborators

Valentina Izrailevitch

Leonie Weissweiler

Lorenzo Xiao

Hinrich Schütze