Evaluating Neural Language Models as Cognitive Models of Language Acquisition

Héctor Javier Vázquez Martínez[1] · Annika Heuser[1]
Charles Yang[1] · Jordan Kodner[2]
[1] University of Pennsylvania · [2] Stony Brook University

GenBench 2023
Singapore

LMs as Cognitive Models of Language

Significant amount of work in this area over the last several years

- Do LMs induce "human-like" (i.e., hierarchical) syntactic representations?
  Yes/Probably: e.g., Gulordava et al. (2018), Papadimitriou et al. (2021), etc.
  No/Probably not: e.g., Chowdhury & Zamparelli (2018), McCoy et al. (2020), etc.
- More recently, are LMs "human-like" models for language acquisition?
  e.g., Huebner et al. (2021), Warstadt & Bowman (2022), etc.

LMs as Cognitive Models of Language

Significant amount of work in this area over the last several years

- Do LMs induce "human-like" (i.e., hierarchical) syntactic representations?
  Yes/Probably: e.g., Gulordava et al. (2018), Papadimitriou et al. (2021), etc.
  No/Probably not: e.g., Chowdhury & Zamparelli (2018), McCoy et al. (2020), etc.
- More recently, are LMs "human-like" models for language acquisition?
  e.g., Huebner et al. (2021), Warstadt & Bowman (2022), etc.

Behavioral Probes and Template-Based Evaluation

- "If it looks like a duck, swims like a duck, and quacks like a duck, then it's a duck."
  Assumes task can only be solved by human-like strategies (Guest & Martin 2023)

LMs as Cognitive Models of Language

Significant amount of work in this area over the last several years

- Do LMs induce "human-like" (i.e., hierarchical) syntactic representations?
  Yes/Probably: e.g., Gulordava et al. (2018), Papadimitriou et al. (2021), etc.
  No/Probably not: e.g., Chowdhury & Zamparelli (2018), McCoy et al. (2020), etc.
- More recently, are LMs "human-like" models for language acquisition?
  e.g., Huebner et al. (2021), Warstadt & Bowman (2022), etc.

Behavioral Probes and Template-Based Evaluation

- "If it looks like a duck, swims like a duck, and quacks like a duck, then it's a duck."
  Assumes task can only be solved by human-like strategies (Guest & Martin 2023)
- Test items are often automatically generated by templates to get around sparsity
  But templates can introduce unintended statistical regularities exploitable by LMs
  And may nevertheless lack variety → lack of empirical coverage of interesting patterns

Two Representative Benchmark Data Sets

1. Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020)

Makes up part of the ongoing CMCL-CoNLL 2023 BabyLM Challenge evaluation sets

- Pairs of grammatical/ungrammatical sentences covering 12 linguistic phenomena
- Automatically created with templates so that the two sentences are minimally distinct
- LM M succeeds on a sentence pair (s_{gram}, s_{ungram}) iff $P_M(s_{gram}) > P_M(s_{ungram})$

Sample Sentence Pair from BLiMP's `adjunct_island` Phenomenon

Grammatical: `Who should Derek hug after shocking Richard?`

Ungrammatical: `Who should Derek hug Richard after shocking?`

Two Representative Benchmark Data Sets

1. Benchmark of Linguistic Minimal Pairs (BLiMP; Warstadt et al., 2020)

2. Zorro (Huebner et al., 2021) - Explicitly Acquisition-Focused

- Directly inspired by BLiMP and adopts the same format with 11 of BLiMPs 12 phenomena
- Restricts vocabulary in order to test LMs trained only on child-directed speech (CDS) such as AO-CHILDES reformatted from the CHILDES collection of CDS corpora
- Zorro was released with BabyBERTa, a transformer that satisfies these constraints

Sample Sentence Pair from Zorro's `local_attractor-in_question_with_aux`

Grammatical: `is the whale getting the person ?`

Ungrammatical: `is the whale gets the person ?`

What do the Probes Actually Test?

Many test pairs are semantically odd or do not test grammaticality at all

- Warstadt & Bowman argue that this is a non-issue since it affects both sentences
- But infelicity affects human judgments of well-formedness in forced choice tasks like what was used to collect judgments for BLiMP (Sprouse et al., 2018)

What do the Probes Actually Test?

Many test pairs are semantically odd or do not test grammaticality at all

- **Warstadt & Bowman argue that this is a non-issue since it affects both sentences**
- **But infelicity affects human judgments of well-formedness in forced choice tasks like what was used to collect judgments for BLiMP (Sprouse et al., 2018)**

Infelicitious example from Zorro: `across_prepositional_phrase`

Grammatical: `the lie on the foot is flat .`

Ungrammatical: `the lie on the foot are flat .`

What do the Probes Actually Test?

Many test pairs are semantically odd or do not test grammaticality at all

- **Warstadt & Bowman argue that this is a non-issue since it affects both sentences**
- **But infelicity affects human judgments of well-formedness in forced choice tasks like what was used to collect judgments for BLiMP (Sprouse et al., 2018)**

Infelicitious example from Zorro: `across_prepositional_phrase`

Grammatical: `the lie on the foot is flat .`

Ungrammatical: `the lie on the foot are flat .`

Hint: `lie` is a noun

What do the Probes Actually Test?

Many test pairs are semantically odd or do not test grammaticality at all

- Warstadt & Bowman argue that this is a non-issue since it affects both sentences
- But infelicity affects human judgments of well-formedness in forced choice tasks like what was used to collect judgments for BLiMP (Sprouse et al., 2018)

Invalid example from BLiMP: ANAPHORA AGREEMENT

Grammatical: `That dancer wouldn't aggravate herself.`

Ungrammatical: `That dancer wouldn't aggravate himself.`

Both are grammatical!

What do the Probes Actually Test?

Many probes do not actually test the intended (or any) structural patterns

They may just rely on linear patterns or even lexical memorization instead

What do the Probes Actually Test?

Many probes do not actually test the intended (or any) structural patterns

They may just rely on linear patterns or even lexical memorization instead

BLiMP Example: ANAPHORA AGREEMENT

only require that the final word in the sentence agrees in number/gender with the first noun

- **The noun and anaphor can be identified with a linear (i.e., non-hierarchical) rule**
- **The mapping between names and conventional gender can only be memorized!**

Grammatical:

`Sherry` `can't forget` `herself`.   `Every` `story` `would disagree with` `itself`.

Ungrammatical:

`Sherry` `can't forget` `himself`.   `Every` `story` `would disagree with` `himself`.

What do the Probes Actually Test?

Many probes do not actually test the intended (or any) structural patterns

They may just rely on linear patterns or even lexical memorization instead

BLiMP Example: SUBJECT-VERB AGREEMENT

only require that the final verb agrees with the first noun

- The noun and verb are adjacent in ⅔ of test sentences
- When a distractor phrase is present, the target noun is still the first noun

→ A linear rule like "the rightmost verb agrees with the leftmost noun" works just fine!

Grammatical:

`Most` <u>`glasses`</u> <u>`scare`</u> `Martin.` `Some` <u>`patients`</u> `who dislike Kendra` <u>`negotiate`</u>`.`

Ungrammatical:

`Most` <u>`glasses`</u> <u>`scares`</u> `Martin.` `Some` <u>`patients`</u> `who dislike Kendra` <u>`negotiates`</u>`.`

What do the Probes Actually Test?

Many probes do not actually test the intended (or any) structural patterns

They may just rely on linear patterns or even lexical memorization instead

BLiMP Example: SUBJECT-VERB AGREEMENT

only require that the final verb agrees with the first noun

- **The noun and verb are adjacent in ⅔ of test sentences**
- **When a distractor phrase is present, the target noun is still the first noun**

→ **A linear rule like "the rightmost verb agrees with the leftmost noun" works just fine!**

We found simple rules like this that achieve 93.97% overall accuracy on Zorro and 84.35% on BLiMP

→ Suggests opportunity for models to "shortcut" these template-based behavioral benchmarks

What do the Probes Actually Test?

Simple handcrafted rules demonstrate that the probes can be shortcutted

Caveat: this is illustrative, not a claim that any given model actually employs a given shortcut. Behavioral probes alone cannot answer this

Zorro	BabyBERTa	Rule
#Sub-Phenoms Rule beats BabyBERTa	—	21/23
Avg Accuracy	78.91%	93.97%
BLiMP	**BabyBERTa**	**Rule**
#Sub-Phenoms Rule beats BabyBERTa	—	61/67
Avg Accuracy	60.72%	84.35%

How complex are these handcrafted rules?

As simple as…

“The 2nd word is `the`” - 100% accuracy on Zorro `wh_question_object`

“Does not start with `Wh`” - 100% on BLiMP `left_branch_island_echo_question`

How complex are these handcrafted rules?

As simple as…

"The 2nd word is `the`" - 100% accuracy on Zorro `wh_question_object`

"Does not start with `Wh`" - 100% on BliMP `left_branch_island_echo_question`

As complex as…

"Word following `had` ends in `n` or there's no word ending in `n`"

Achieves 88.40% on Zorro `irregular_verb`

"Last word ends in `s` and (either first word is any of `{Many, These, All, Most, Those}` OR the 2nd word is `lot`) OR the 2nd word ends in `s`"

Achieves 71.35% on BLiMP `principle_A_c_command`

Revisiting an *N*-Gram Baseline

Linear 5-Gram models over words or tags perform well

- Especially compared to BabyBERTa (Huebner et al, 2021) trained on AO-CHILDES
- Sub-phenomena solvable by an *n*-gram model are irrelevant for the task:
  If something like an *n*-gram model or simple rule can solve these, we can't conclude anything either way about structural knowledge from them

Zorro	BabyBERTa	5-Gram Word	5-Gram Tag	Either 5-Gram
#Sub-Phenoms 5-Gram beats BabyBERTa	—	8/23	8/23	11/23
Avg Accuracy	78.91%	63.44%	57.59%	—
BLiMP	**BabyBERTa**	**5-Gram Word**	**5-Gram Tag**	**Either 5-Gram**
#Sub-Phenoms 5-Gram beats BabyBERTa	—	18/67	10/67	23/67
Avg Accuracy	60.72%	50.72%	37.93%	—

Revisiting an *N*-Gram Baseline

Linear 5-Gram models over words or tags perform well

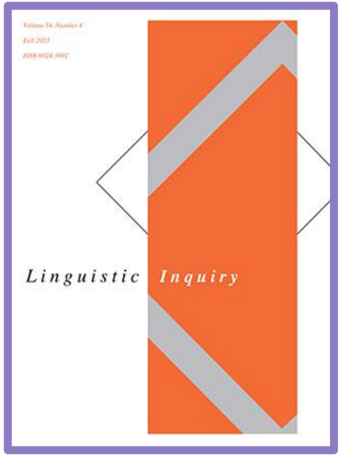
- Especially compared to BabyBERTa (Huebner et al, 2021) trained on AO-CHILDES
- Sub-phenomena solvable by an *n*-gram model are irrelevant for the task:
  If something like an *n*-gram model or simple rule can solve these, we can't conclude anything either way about structural knowledge from them

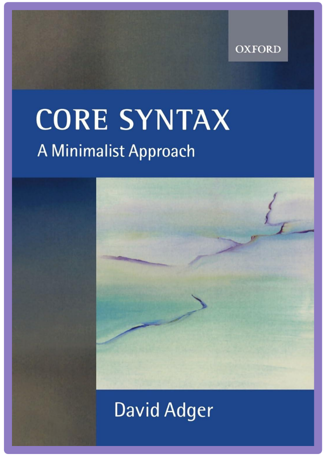
→ Both the data set and the "duck test" logic are off the mark 🦆

- Much of the data cannot distinguish linear from hierarchical representations
- Much of the probes that requires hierarchical representations in principle could be short-cutted in practice

Proof-of-Concept: The LI-Adger Dataset

- Collection of **“real” test sentences** from *Linguistic Inquiry* journal and *Core Syntax* textbook (collected by Sprouse et al., 2013)
- **Several human judgments** available for each
- Linguistic **theory-agnostic** empirical phenomena
- **Reliable, replicable and statistically powerful** (Sprouse & Almeida, 2012; Sprouse et al., 2013; Sprouse & Almeida 2017; Sprouse et al. 2017; among others)





Proof-of-Concept: The LI-Adger Dataset

Comprehensive coverage of linguistic phenomena

- Avoids the template bias problem

Sentences manually created by experts

- Control for semantic implausibility

Magnitude Estimation judgements

- Allow for comparisons *across* minimal pairs
- Contra Forced-Choice, which treats sentence acceptability as a categorical measure

Multiple judgements per sentence

- Allows correlation with human judgments
- And between-human and between-model judgments

Overview of the Benchmarks

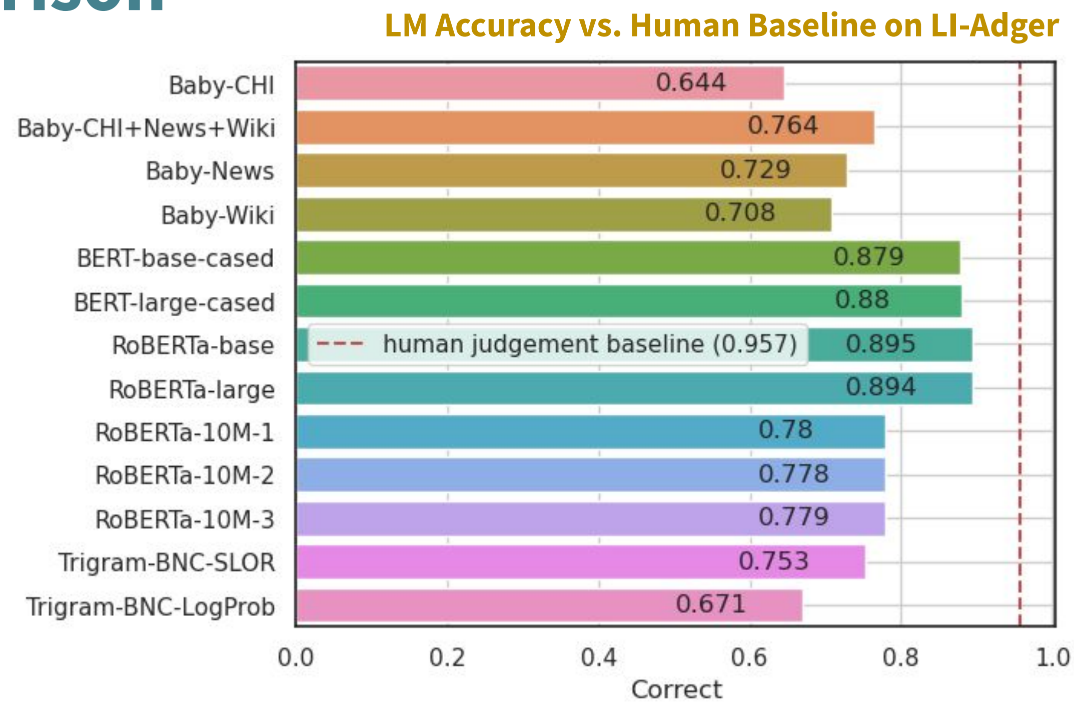
Property	BLiMP	Zorro	LI	Adger
			LI-Adger	
Source	Templates	Templates	*Linguistic Inquiry* 2001-10	*Core Syntax* (Adger, 2003)
Semantic Implausibility	Yes	Yes	No	No
#Sub-Phenoms (paradigms)	67	23	150	105
#Min. Pairs per Sub-Phenom	1000	2000	8	8
Human Judgements	FC	(None)	LS, <u>ME</u>, FC	<u>ME</u>, FC
#Judgements per Sentence	< 1	N/A	13 (ME)	10 (ME)

FC - Forced Choice
LS - Likert Scale
ME - Magnitude Estimation

More Model Comparison

Human baseline when evaluated against categorical expert labels is much higher (0.957) than in BLiMP (0.886)

The trigram model matches the performance of all models trained on a "developmentally plausible" amount of data.

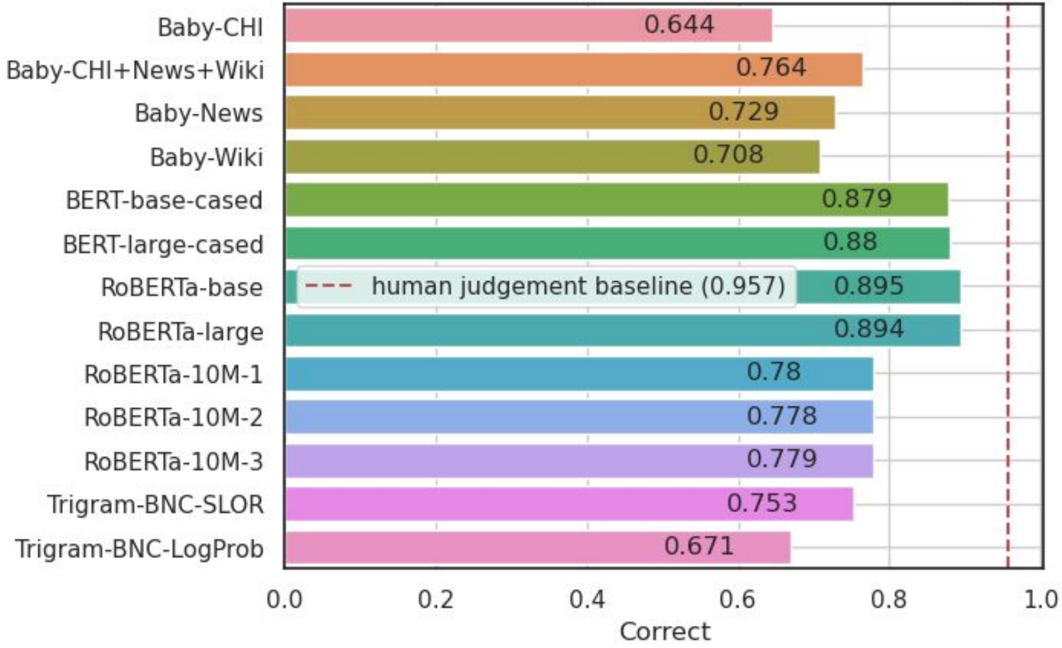


More Model Comparison

Human baseline when evaluated against categorical expert labels is much higher (0.957) than in BLiMP (0.886)

The trigram model matches the performance of all models trained on a “developmentally plausible” amount of data.

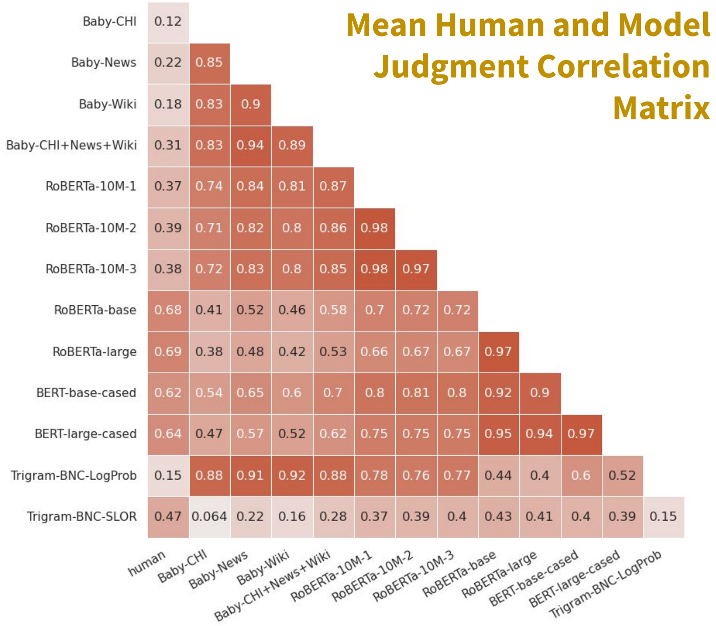
LM Accuracy vs. Human Baseline on LI-Adger

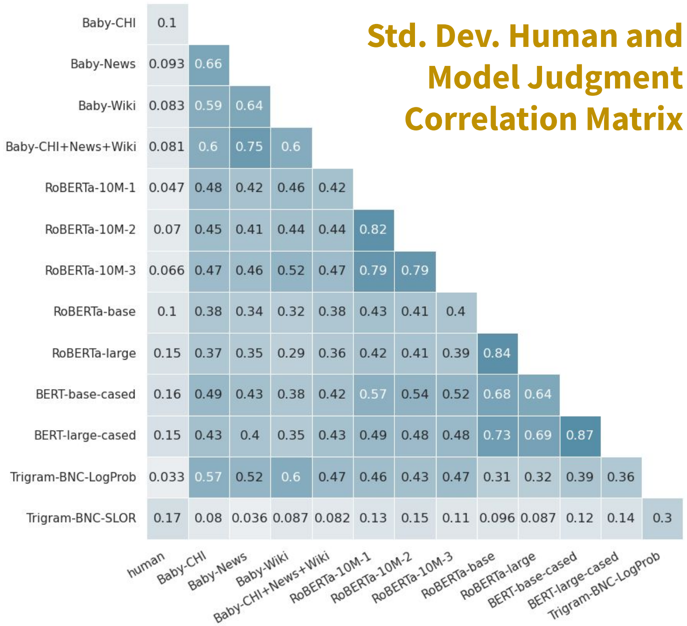


Our finding:

Co-occurrence statistics may yield high performance, but how a model’s judgements behave across structures may reveal a more human-like strategy.

Proof-of-Concept: The LI-Adger Data Set





Our finding:

Even when LMs achieve good accuracy, they usually correlate better with each other than with humans. The LMs trained on CHILDES correlate the most weakly.

Main Conclusions

Template-based behavioral benchmarks have serious weaknesses

- They contain non-hierarchical shortcuts that LMs may exploit
- Sentences are insufficiently varied and may be unnatural

→ Success on such benchmarks may tell us little about whether LMs are plausible cognitive models for language acquisition or otherwise

→ When using behavioral benchmarks, we recommend using more naturalistic sentences with many human judgments like the LI-Adger data set

Our code is available on Github.
See our paper for more information!

Penn
UNIVERSITY of PENNSYLVANIA

Stony Brook
University