Linguistics Lab

13 March 2024 • NLU

On NLP/ML/AI/etc and Linguistics

Lukewarm take: NLP and Linguistics are not the most compatible fields

= linguistics + ML people ask different questions, don't work well together

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

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Lol, lmao even 🙄

Hot (?) take: This is a contingent fact, not a necessary one

Why think about "Language"/Linguistics (in NLP)?

Linguistics can give useful insight!

- "Every time I fire a linguist, the performance of the speech recognizer goes up" (Frederick Jelinek, NLP pioneer)
- Counterpoint: RNNs! (Jeffrey Elman, *Finding Linguistic Structure in Time*)
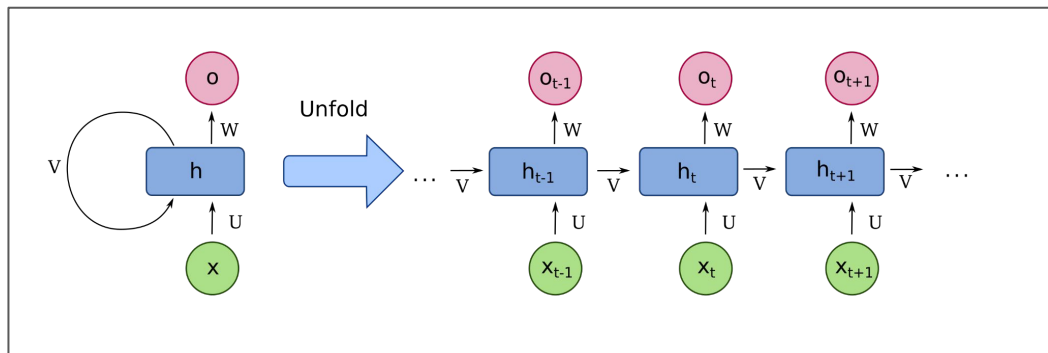
Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu[*1] and Tri Dao[*2]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Computer Science, Princeton University
agu@cs.cmu.edu, tri@tridao.me

Google DeepMind   2024-3-1

Griffin: Mixing Gated Linear Recurrences with Local Attention for Efficient Language Models

Soham De[*1], Samuel L. Smith[*1], Anushan Fernando[*1], Aleksandar Botev[*1], George Cristian-Muraru[*1], Albert Gu[2], Ruba Haroun[1], Leonard Berrada[1], Yutian Chen[1], Srivatsan Srinivasan[1], Guillaume Desjardins[1], Arnaud Doucet[1], David Budden[1], Yee Whye Teh[1], Razvan Pascanu[1], Nando De Freitas[1] and Caglar Gulcehre[1]
[*]Equal contributions, [1]Google DeepMind, [2]Work done while at Google DeepMind

Language raises interesting problems

Dialect prejudice predicts AI decisions about people's character, employability, and criminality

Valentin Hofmann[1-3*†], Pratyusha Ria Kalluri[4], Dan Jurafsky[4], Sharese King[5*]

[1]Allen Institute for AI   [2]University of Oxford   [3]LMU Munich
[4]Stanford University   [5]The University of Chicago

How do you train a (trillion?) parameter LM to have good performance on low-resource languages?

How does linguistic variation affect model responses? How *should* it?
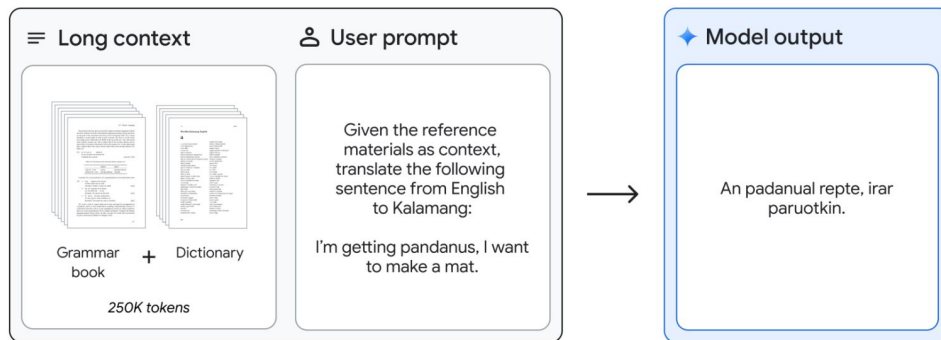
LlMs ArE jUsT sToChAsTiC pArRoTs

Figure 3 | Given a reference grammar book and a bilingual wordlist (dictionary), Gemini 1.5 Pro is able to translate from English to Kalamang with similar quality to a human who learned from the same materials.

Kalamang (200 native speakers, ~no digital training data)

Questions I'm interested in

Recall Poverty of the Stimulus argument from lecture:

Humans make consistent generalizations from very minimal, ambiguous data

=> Humans have innate bias towards "language" (whatever that looks like)

Two interesting points:

- Humans systematically choose A over B
- Humans learn to choose A over B with *very little data*

Q1: How do we get LLMs to make similar generalizations?

Suppose you want to turn natural language descriptions into code (method calls, etc).

How can you ensure that an LLM generalizes to:

- Arbitrarily deep function composition?
- Arbitrary (correct) variable identity?

These properties encapsulate *compositionality*

- This is a DEFINING property of natural language in general

Compositionality

jump
JUMP

turn around left
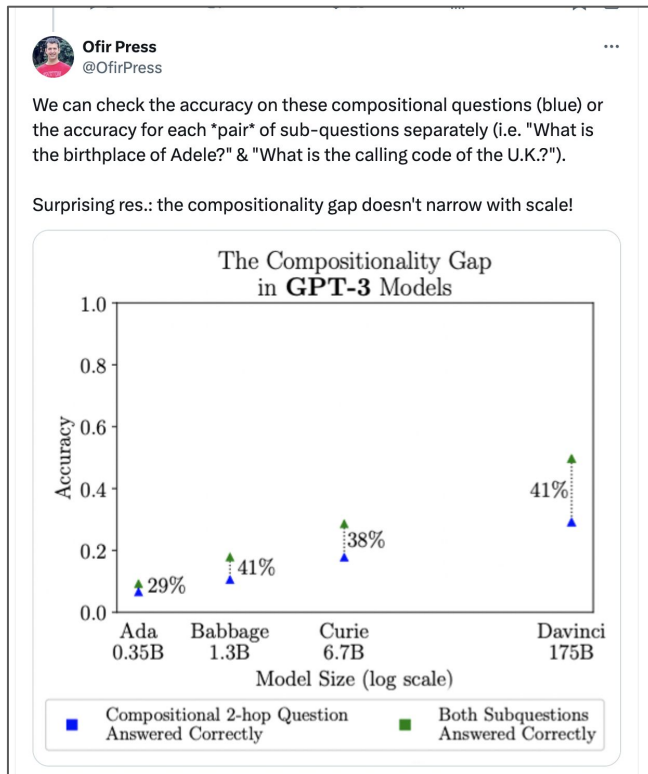LTURN LTURN LTURN LTURN

jump thrice and turn left twice
JUMP JUMP JUMP LTURN LTURN

jump opposite left after walk twice
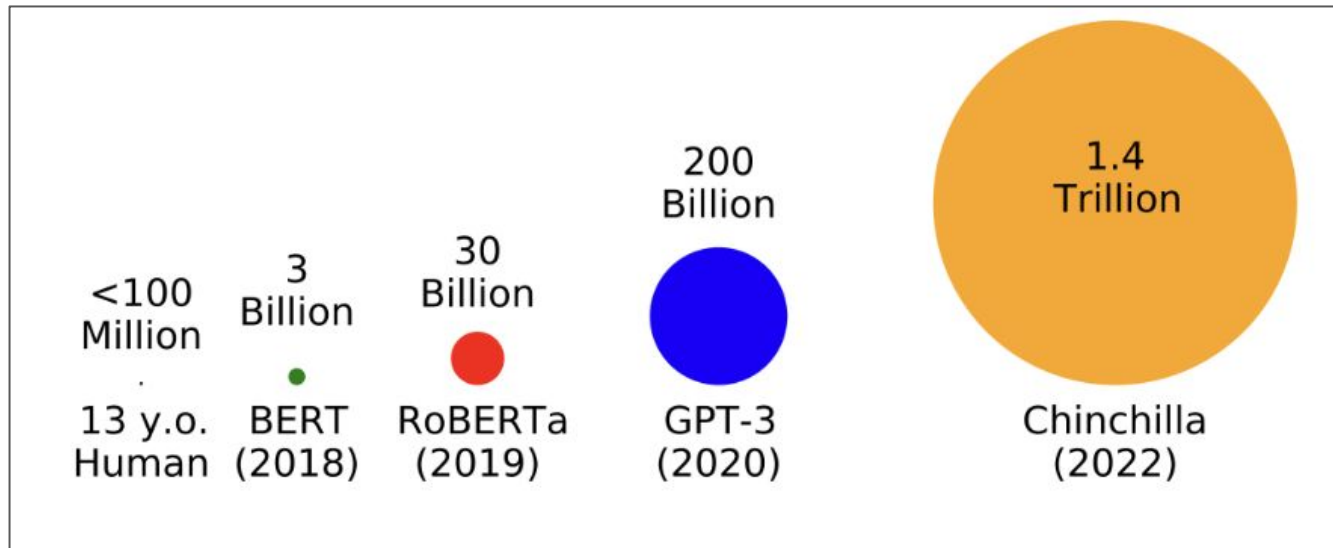WALK WALK LTURN LTURN JUMP

Figure 1: SCAN maps commands to actions

Small scale

Ofir Press
@OfirPress

We can check the accuracy on these compositional questions (blue) or the accuracy for each *pair* of sub-questions separately (i.e. "What is the birthplace of Adele?" & "What is the calling code of the U.K.?").

Surprising res.: the compositionality gap doesn't narrow with scale!

The Compositionality Gap in **GPT-3** Models

Accuracy / Model Size (log scale)

Ada 0.35B — 29%
Babbage 1.3B — 41%
Curie 6.7B — 38%
Davinci 175B — 41%

■ Compositional 2-hop Question Answered Correctly
■ Both Subquestions Answered Correctly

Large scale

Q2: How do we get LLMs to be as efficient as people?

200
Billion

30
Billion

1.4
Trillion

3
Billion

<100
Million

13 y.o.   BERT   RoBERTa   GPT-3   Chinchilla
Human   (2018)   (2019)   (2020)   (2022)

BabyLM challenge: train a language model on human-realistic data