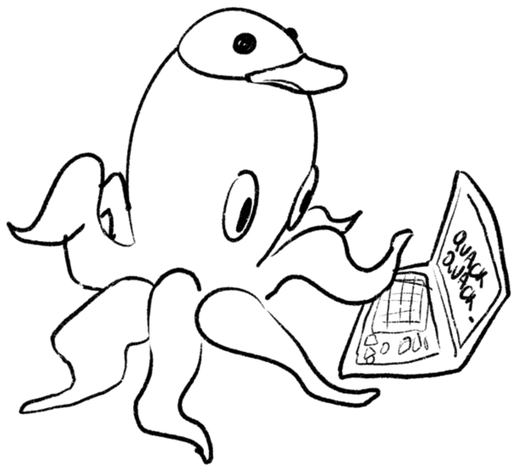
Language Models Understand Us, Poorly

Jared Moore

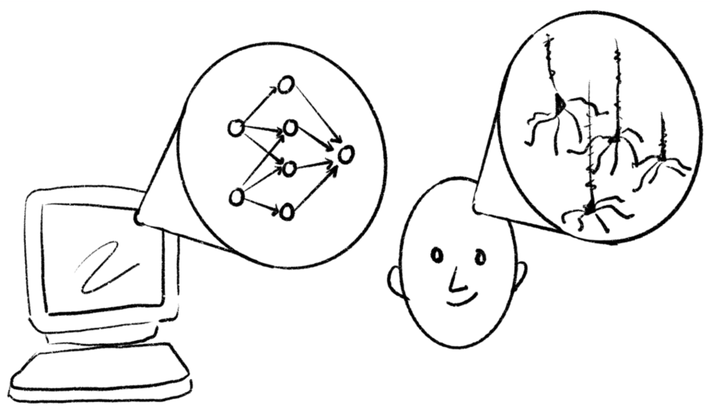
Paper: bit.ly/understand_poorly
Slides: bit.ly/understand_poorly_slides

Three Views on Understanding



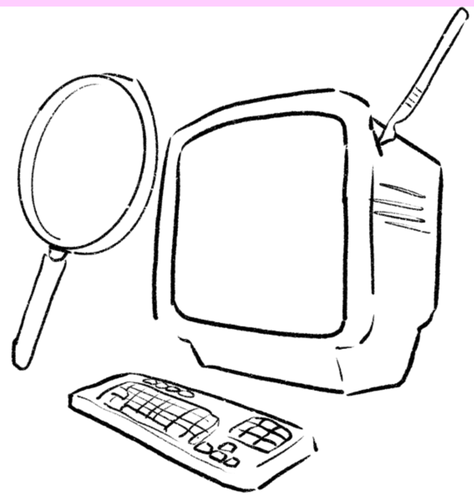
Understanding-as-**reliability**

- There is no distinction between human and machine understanding.
- Models will close that gap soon (Agüera y Arcas 2022).
- Scale is paramount (Chowdhery et al. 2022).



Understanding-as-**mapping**

- There is a "barrier of meaning" which separates human from machine understanding (Bender et al. 2021).
- Syntax is separate from semantics.



Understanding-as-**representation**

There is a continuum of understanding…

- but it depends on demonstrating the same skills.
- (Language models have a "sorta" comprehension; they perform well in some domains (Dennett 2017).)

Motivation

A recent meta-survey of NLP researchers (Michael et al. 2022) found that a mean of…

- 51% thought LMs understand language
- 67% thought multimodal models understand language
- And 36% thought text-only evaluation could measure language understanding.

What do models understand? *(What do we understand?)*

How to climb the right hill?

	Necessary	Not Necessary
Sufficient		Understanding-as-**representation**
Not Sufficient	Understanding-as-**reliability**	Understanding-as-**mapping**

Humans assume a similarity of representation. (Remember Eliza?)
•We can't make that assumption with our models. (cf. Michael, 2020)

Under-specification

(Failures of assuming a similarity of representation)

Uni-modal underspecifications, e.g.

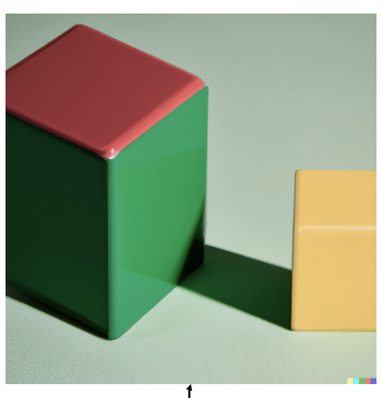
•Entailments
- If the artist slept, the actor ran. Yes or no, did the artist sleep?

•Copying style and answering
- t.w.o.p.l.u.s.t.w.o.e.q.u.a.l.s.w.h.a.t.?

•Long context window; truthiness

Multi-modal underspecifications…

- Are no better than chance (Thrush et al. 2022).
- E.g. DALL-E "A red cube, on top of a yellow cube, to the left of a green cube" [link]



Toward a Similarity of Representation

(How to correct models' inductive biases)

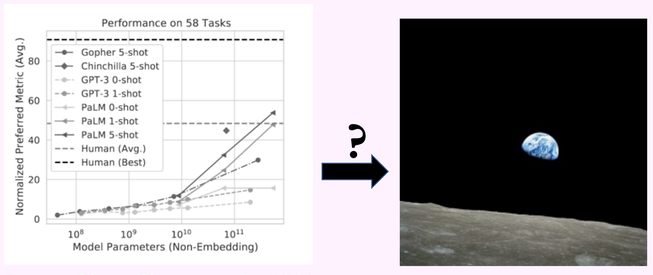
Add Social domains

Models are only slightly better than chance at theory of mind (Sap et al. 2022).
•And we're only starting to see good tests for the components of moral reasoning (Weidinger, Reinecke, and Haas 2022).

Increase Generalization

•By 5yo, the average American child has heard between 10 and 50 million words (Sperry, Sperry, and Miller 2019).
•Embodiment is needed eventually (Lynott et al. 2020; Bisk et al. 2020).

Is Scale Enough?



From Chowdhery et al. (2022)

- Models see 10-100,000 times more words than a kid
- E.g."for PaLM**, data begins to repeat** in some of our subcorpora after 780B tokens" (Chowdhery et al. 2022) (emphasis added)

Sorta Understands != Understands

•"Computers which understand" is most often false advertising
- *Sometimes* it is a statement about theoretical AI

•In a theoretical sense, LMs may understand us poorly but in a pragmatic sense they do not understand us at all.

How to have a more pragmatic NLP

Probe model internals.
•Black box and white box them

Add more of human language.
•E.g. intersubjective, multi-agent environments.
•CHILDES database of childhood language learning (MacWhinney 2000; Linzen 2020).

Measure what models can learn.
•E.g. how many different streams of data (or "world scopes" (Bisk et al. 2020) must we add to models to make them more reliable?

Works Cited

- Agüera y Arcas, Blaise. 2022. "Do Large Language Models Understand Us?" Dædalus 151 (2). https://doi.org/10.1162/DAED_a_01909.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.
- Bisk, Yonatan, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, et al. 2020. "Experience Grounds Language." arXiv:2004.10151 [Cs], November. http://arxiv.org/abs/2004.10151.
- Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, et al. 2022. "PaLM: Scaling Language Modeling with Pathways." arXiv:2204.02311 [Cs], April. http://arxiv.org/abs/2204.02311.
- Dennett, Daniel C. 2017. From Bacteria to Bach and Back: The Evolution of Minds. WW Norton & Company.
- Linzen, Tal. 2020. "How Can We Accelerate Progress Towards Human-Like Linguistic Generalization?" In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5210–17. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.465.
- Lynott, Dermot, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. "The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words." Behavior Research Methods 52 (3): 1271–91.
- MacWhinney, Brian. 2000. The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates.
- Michael, Julian. 2020. "To Dissect an Octopus: Making Sense of the Form/Meaning Debate." Julian Michael. https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html.
- Michael, Julian, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, et al. 2022. "What Do NLP Researchers Believe? Results of the NLP Community Metasurvey." arXiv. https://doi.org/10.48550/arXiv.2208.12852.
- Sap, Maarten, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. "Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs." arXiv. http://arxiv.org/abs/2210.13312.
- Sperry, Douglas E., Linda L. Sperry, and Peggy J. Miller. 2019. "Reexamining the Verbal Environments of Children From Different Socioeconomic Backgrounds." Child Development 90 (4): 1303–18. https://doi.org/10.1111/cdev.13072.
- Thrush, Tristan, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality." arXiv:2204.03162 [Cs], April. http://arxiv.org/abs/2204.03162.
- Weidinger, Laura, Madeline G. Reinecke, and Julia Haas. 2022. "Artificial Moral Cognition: Learning from Developmental Psychology." Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/tnf4e.