Theory of Mind: Children aged 5-12 vs. a Large Language Model (GPT3)

Max van Duijn, Werner de Valk, Bram van Dijk, & Peter van der Putten

Keywords
Theory of Mind, Large Language Models, social cognition, artificial intelligence

Objectives
Large Language Models (LLMs) such as OpenAI's ChatGPT have recently shown striking abilities to produce natural language, performing at or beyond human level on many tasks including conversation and creative writing. Here we feed language-based Theory of Mind (ToM) tests to GPT3 [1] and compare its performance to that of children aged 5-12. Our aim is not to test whether LLMs like GPT3 possess ToM. Rather, comparing the model output with that of children of different ages allows us to map out parallels in commonly made mistakes, and to identify conditions that influence performance.

Methods
As part of a larger study* we ran a test battery including the Strange Stories Task [2] for children aged 5-7 (n=44) and a test of recursive ToM [3] for children aged 7-12 (n=42). Tests were presented verbally (voice-over) and in written form, accompanied by illustrations, and questions were multiple choice and/or open text fields eliciting brief motivations. We wrote a Python script feeding the same questions to the text-davinci-003 model via the OpenAI API, with minimal adaptations to prompt the desired output format.

Results
Parallel to children in our sample, GPT3 performs quite well overall, but makes more errors on more complex questions, particularly those involving irony, sarcasm, and recursive ToM. Also like humans [4], the model is highly sensitive to how information regarding these more challenging ToM questions is being presented.

Conclusions
Our results are in line with existing findings on ToM in children [5] and adults [3], and deepen current understanding of the limitations of LLMs [6-8]: they falter when a grounded understanding of the (in this case social) world is required. We discuss the implications and future potential for research into the mechanisms that support ToM, and into what is needed for a ToM capacity to evolve.

References
* Ethics: *Anonymised University* ID 2020-1-02
[1] Tom B. Brown, Benjamin Mann, Nick Ryder et al. 2020. 'Language Models are Few-Shot Learners'. arXiv:10.48550/ARXIV.2005.14165
[2] Francesca G.E. Happé. 1994. 'An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults'. *J Autism Dev Disord* 24: 129-154.
[3] adapted from Penny A. Lewis, R. Rezaie, R. Brown, N. Roberts, & R.I.M. Dunbar. 2011. 'Ventromedial prefrontal volume predicts understanding of others and social network size.' *NeuroImage* 57:(4): 1624-1629.

[4] Max J. van Duijn, I. Sluiter, & A. Verhagen. 2015. ‘When narrative takes over: The representation of embedded mindstates in Shakespeare’s Othello’. *Language and Literature* 24:(2): 148–166.
[5] A. Nicolopoulou & B. Ünlütabak. 2017. ‘Narrativity and mindreading revisited. Children’s understanding of theory of mind in a storybook and in standard false belief tasks’. In Ketrez, F. et al. (Eds). Social environment and cognition in language development. studies in honor of Ayhan Aksu-Koç (pp. 151-166). Amsterdam: Benjamins.
[6] Maarten Sap, Ronan Le Bras, Daniel Fried, & Yejin Choi. 2022. ‘Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs’. arXiv:arXiv:2210.13312v1
[7] Jaan Aru, Aqeel Labash, Oriol Corcoll, & Raul Vicente. 2022. ‘Mind the Gap. Challenges of Deep Learning Approaches to Theory of Mind’. arXiv:2203.16540v2
[8] Anna Rogers, Olga Kovaleva, & Anna Rumshisky. 2020. ‘A Primer in BERTology: What We Know About How BERT Works’. arXiv:2002.12327v3