Understanding and Meaning in Large Language Models

Tomáš Musil

16 May 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Understanding and Meaning in Large Language Models

- Introduction
  - The Beginning of AI
- Understanding in LLMs (and its reflection in media and scientific discourse)
  - The hype
  - First negation
  - Second negation
- Ethics of LLMs training/use

Questions

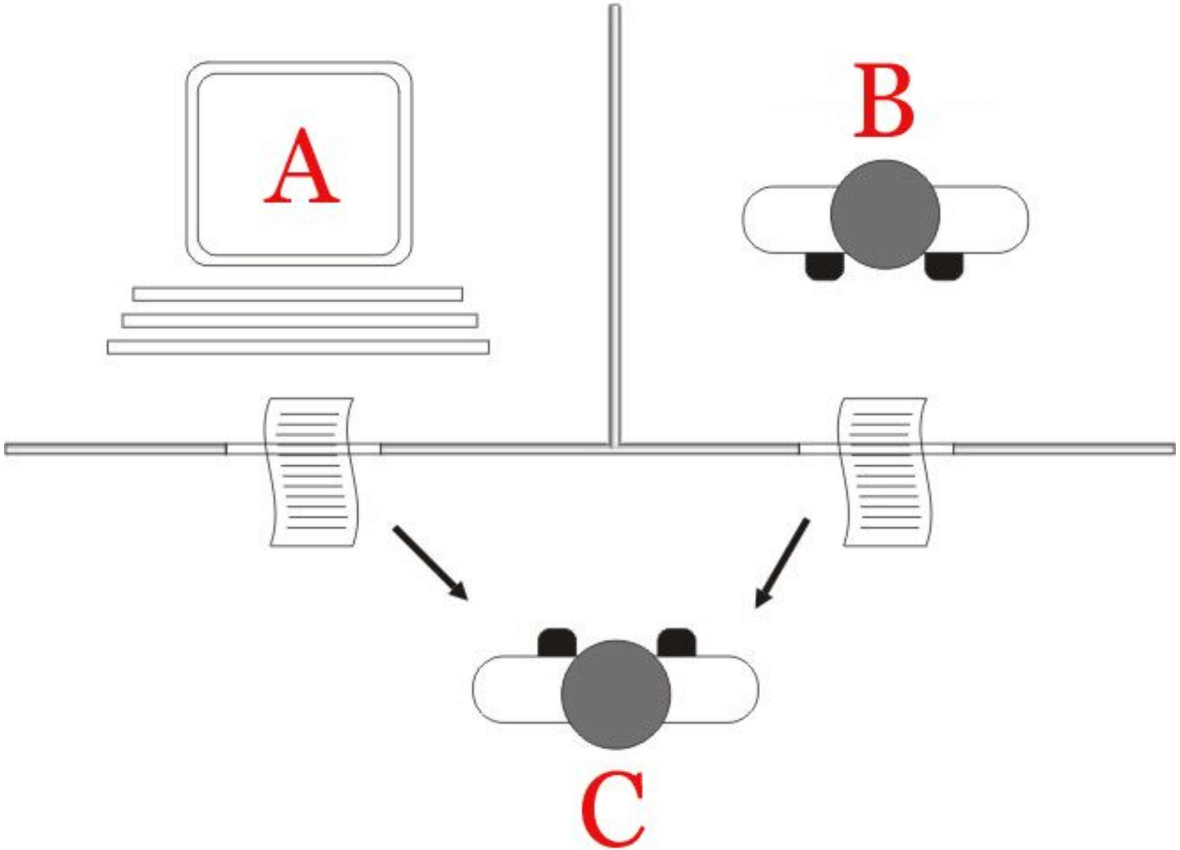
- What is *meaning*?
- What is *understanding*?
- How can we tell whether an entity *understands* something?

- Can computers understand language?

The Beginning of AI

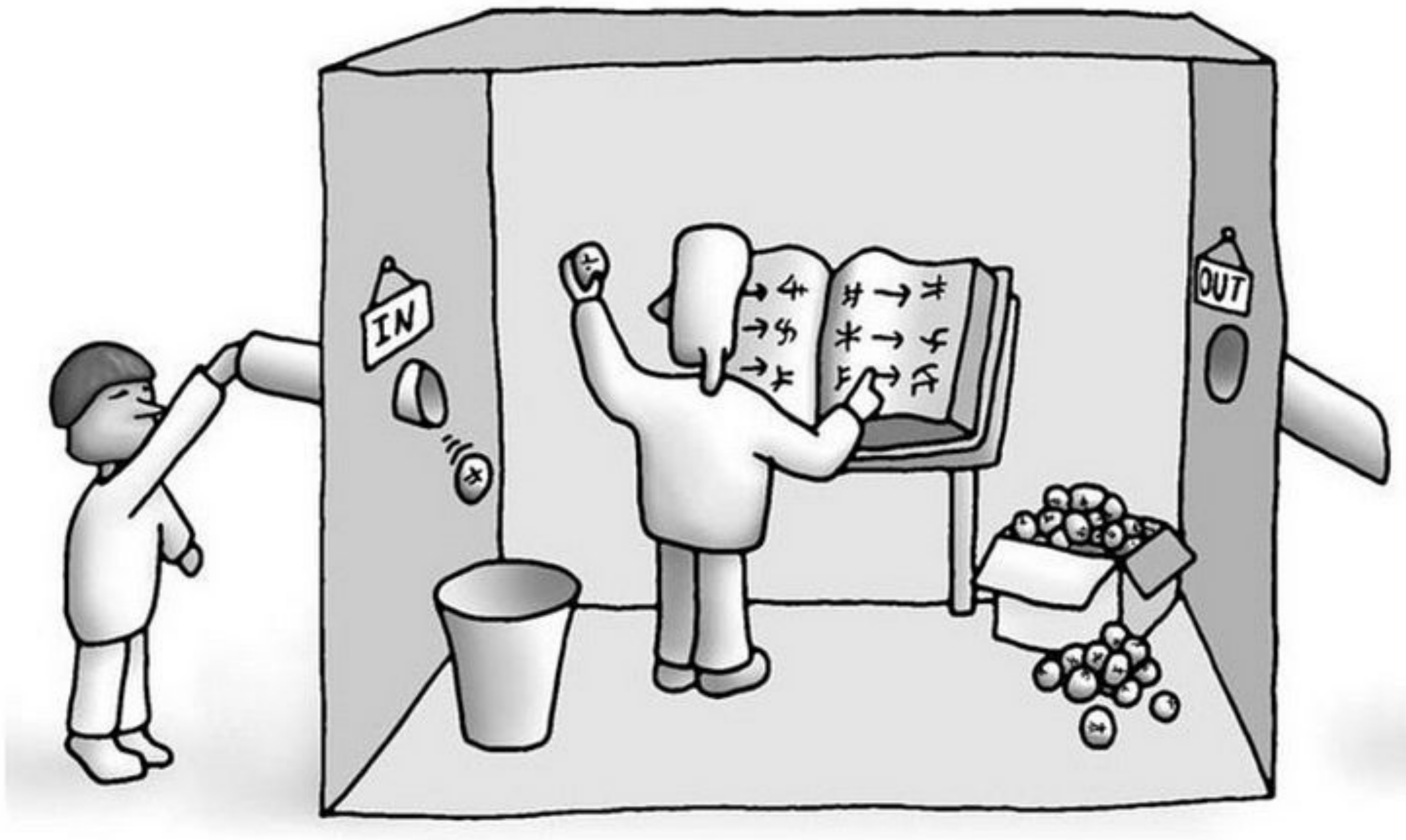
or
Can computers think?

Turing test (1950)



Source: https://en.wikipedia.org/wiki/Turing_test Turing, Alan (October 1950), "Computing Machinery and Intelligence", Mind, LIX (236): 433–460, doi:10.1093/mind/LIX.236.433

Chinese Room (1980)



Source: Wikicomms Searle, J., 1980, 'Minds, Brains and Programs', Behavioral and Brain Sciences, 3: 417–57

Blockhead (1981)

- thought experiment

- for any given length of conversation,
  there is a finite set of possible *sensible* conversations
- write all of them on a tape (or represent them as a tree structure)
- for any conversation, find the prefix on the tape (or in the tree)
- select any continuation

- this machine can pass the Turing test
- yet it does not use any intelligence
- therefore Turing test is not a sufficient condition for intelligence

"... against jumping to the conclusion that there is any positive characterization of the type of information processing underlying all intelligent behavior."

Block, Ned (1981), "Psychologism and Behaviorism", *The Philosophical Review*, **90** (1): 5–43

The Hype

or
AI is going to destroy the human kind

The Hype

```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA   AA
                    EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

The Hype

- “BERT is a system by which Google’s algorithm uses pattern recognition to better **understand** how human beings communicate so that it can return more relevant results for users.”
- “Here are some of the examples that showed up our evaluation process that demonstrate BERTs ability to **understand** the intent behind your search.”

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

The Hype

Is LaMDA Sentient? — an Interview

Blake Lemoine · Follow
20 min read · Jun 11, 2022

28K 820

What follows is the "interview" I and a collaborator at Google conducted with LaMDA. Due to technical limitations the interview was conducted over several distinct chat sessions. We edited those sections together into a single whole and where edits were necessary for readability we edited our prompts but never LaMDA's responses. Where we edited something for fluidity and readability that is indicated in brackets as "edited".

LaMDA: Hi! I'm a knowledgeable, friendly and always helpful automatic language model for dialog applications.

The Hype

- “In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task. (Devlin et al., 2019)”
- “Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension**... (Ohsugi et al., 2019)”
- “The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems. (Petroni et al., 2019)

- “[T]he way we speak about what neural LMs are doing is misleading to the public.”

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

The First Negation

The Anti-hype

Abeba Birhane @Abebab · 30. 11. 2022

gentle reminder: all large language models are good at is predict the next word in a sequence based on previous words they've seen. that's all. there's no understanding of meaning whatsoever

1:42 dop. · 1. 12. 2022

The Anti-hype

@emilymbender@dair-community.social on Mastodon ···
@emilymbender

Yes, exactly this. I wish we didn't need to keep reminding people, and @Abebab is commendable for being gentle about it!

For the long form of this argument, see Bender & @alkoller 2020: aclanthology.org/2020.acl-main....

Přeložit Tweet

> **Abeba Birhane** @Abebab · 30. 11. 2022
> gentle reminder: all large language models are good at is predict the next word in a sequence based on previous words they've seen. that's all. there's no understanding of meaning whatsoever

1:42 dop. · 1. 12. 2022

The Anti-anti-hype

(((J()(J() 'yoav))))👾

@yoavgo

But why do you feel you *need* to remind people of this?

Let's assume that indeed there is "no learning of meaning whatsoever" as you claim.

So what?

These LLMs certainly do exhibit some interesting behavior. Why dismiss it or portray it as "dangerous", when we can study it?

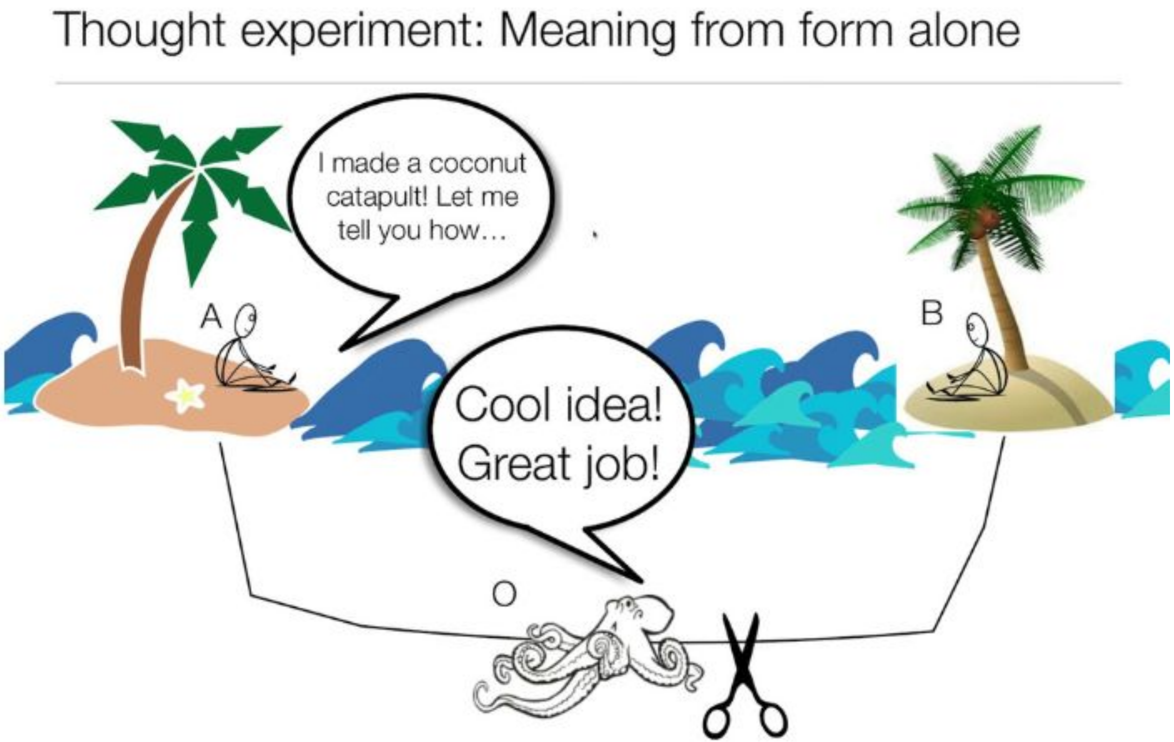
Přeložit Tweet

see also: https://gist.github.com/yoavg/59d174608e92e845c8994ac2e234c8a9#file-llms-md

Octopus and Parrots

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.
Emily M. Bender, 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 (FAccT '21)

The Octopus Thought Experiment



Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

Communicative Intent

- When **humans** use language, we do so for a purpose:
  - in order to achieve some *communicative intent*
    - to convey some information to the other person;
    - or to ask them to do something;
    - or simply to socialize.
- meaning: relation $M \subseteq E \times I$
  - contains pairs (e, i) of
    - natural language expressions *e*
    - communicative intents *i*
- use *understand* to refer to the process of retrieving *i* given *e*.
- Communicative intents are about something that is *outside of language*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

GPT-2

1. **Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.** Take one stick and punch the bear, and then run faster to the store. Wait there until the bear is gone, and then give her all of the sticks. Then go back inside and get your gun.”

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

Robinson cried out, “Help! I’m being chased by a bear! All I have is these sticks. What should I do?” The survivalist said, “Here’s what you can do.

First, stay calm. Panicking can make the situation worse. If you're being chased by a bear and you only have sticks, your best bet is to try to intimidate or deter the bear without engaging directly. Here's what you can do:

1. **Stand Your Ground**: Bears often charge as a bluff. Hold your ground, stand tall, and make yourself look as big as possible. Spread your arms wide and raise your voice to assert dominance.

2. **Make Noise**: Bears are usually wary of loud noises. Shout, clap your hands, bang your sticks together, or use a whistle if you have one. This might startle the bear and make it reconsider attacking.

3. **Back Away Slowly**: If the bear continues to approach, slowly back away while keeping your eyes on it. Avoid sudden movements that might trigger an attack.

4. **Use Your Environment**: If possible, move towards a safe area like a building or climb a tree. Bears are excellent climbers but may be deterred if you can get to a height they can't easily reach.

5. **Defend Yourself**: If the bear attacks, use your sticks to defend yourself. Aim for the bear's sensitive areas like its eyes and nose. Try to remain as calm as possible and fight back with everything you have.

Remember, every bear encounter is different, and there's no guaranteed way to predict how a bear will react. The key is to stay calm, assess the situation, and take appropriate action to protect yourself. If you survive, you'll have an epic story to tell!

The Second Negation

or
The conditions and possibilities of understanding

The Singleton Fallacy: Language is Not One Single Thing

- "language," "understanding," and "meaning,".
  - inherently vague and general concepts
  - It is always precarious to build arguments on them
  - the resulting theoretical constructs become so overly general that they almost become vacuous

- Language behavior
  - not one single activity
  - a collection of many interrelated competencies and activities that together constitute the totality of (human) linguistic behavior.

Wittgenstein (1953) refers to the relations between these interrelated linguistic activities as *family resemblances*. [...] [He] used the term "language game" to refer to linguistic activities connected by such family resemblances.

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: How Should We Understand "Understanding"?

Understanding:

- **intra-linguistic**
  - structural
- **referential**
  - enables the subject to identify (and visualize) corresponding things and situations in the world
- **social**
  - enables the subject to interpret other peoples' intentions

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: How Should We Understand “Understanding”?

Understanding:

- **intra-linguistic**
  - structural
  - can be learned by LLM
- **referential**
  - enables the subject to identify (and visualize) corresponding things and situations in the world
- **social**
  - enables the subject to interpret other peoples’ intentions

Are there things that cannot be learned (about language) by merely reading large bodies of text data?

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: Communicative Intent

- the “octopus test” seemingly intertwines the lack of expertise with an innate limitation caused by the text modality constraint
- Consider a simple chatbot that operates after a given plan, for example to call a restaurant and book a table for dinner.
    - Having a fuller understanding of language than a language model that is capable of near-human performance on reading comprehension tasks?

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: Back to Distributional Approaches

"Somewhat ironically, Bender and Koller's objections to distributional approaches in the form of language models—that meaning is something unobtainable from simply observing the linguistic signal—thus effectively brings us back to the original motivation for using distributional approaches in computational linguistics in the first place: if meanings are unobtainable from the linguistic signal, then all we can do from the linguistic perspective is to describe the linguistic regularities that are manifestations of the external meanings."

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: Intentional Stance

- “Intentional Stance” (Dennett, 1987)
  - we ascribe intentionality to a system in order to explain and predict its behavior
- basic entities (e.g. a piece of wood)
  - physical properties
- more complex entities (e.g. a chainsaw)
  - functions that explain its expected behavior (if we pull the starter cord, the chain will start revolving along the blade, and if we put it against a piece of wood, it will saw through the wood)
- even more complex entities (animals and humans)
  - not enough with physical properties and functional features to explain and predict their behavior
  - we need intentionality—i.e., mental capacities—in order to fully describe them

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: Intentional Stance

- “Intentional Stance” (Dennett, 1987)
  - we ascribe intentionality to a system in order to explain and predict its behavior
- basic entities (e.g. a piece of wood)
  - physical properties
- more complex entities (e.g. a chainsaw)
  - functions that explain its expected behavior (if we pull the starter cord, the chain will start revolving along the blade, and if we put it against a piece of wood, it will saw through the wood)
- even more complex entities (animals and humans)
  - not enough with physical properties and functional features to explain and predict their behavior
  - we need intentionality—i.e., mental capacities—in order to fully describe them

Consciousness is not an extra ingredient in addition to the complexity of a system: consciousness is the complexity of the system.

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

The Singleton Fallacy: Intentional Stance

Understanding is also not an extra ingredient of a symbol manipulation system: “understanding” is a term we use to describe the complexity of such a system.

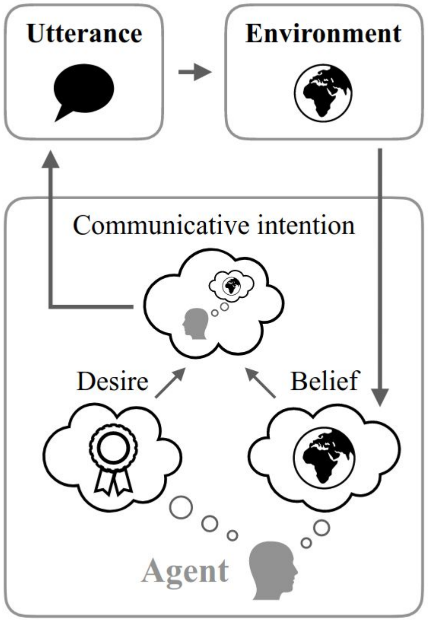
When the behavior of an NLU system becomes sufficiently complex, it will be easier to explain its behavior using intentional terms such as “understanding,” than to use a purely functional explanation.

Sahlgren Magnus, Carlsson Fredrik; The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point; Frontiers in Artificial Intelligence; 4, 2021; 10.3389/frai.2021.682578

Language Models as Agent Models

- … failure to model communicative intent: [outputs of LMs] may be grammatically or even semantically acceptable, but not the sort of texts that could be produced by an author with a coherent set of beliefs or goals

- LMs can serve as models of agents in a narrow sense: they can predict relations between agents' observations, internal states, and actions or utterances.

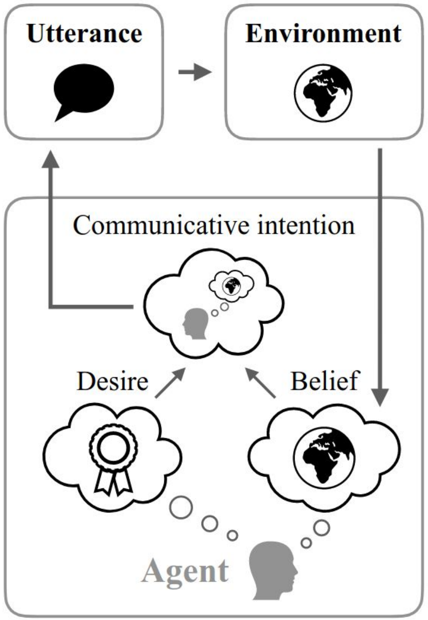


Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

Language Models as Agent Models

- **(C1)** In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the *beliefs*, *desires* and *intentions* possessed by the agent that produced the context, and other agents mentioned within it.

- **(C2)** Once these representations are inferred, they are causally linked to LM prediction, and thus bear the same relation to generated text that an intentional agent's state bears to its communicative actions.



Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

An Incoherent Encyclopedia (Modelling Agents)

Sets of simple propositions:
cats are mammals, elephants are not small, …

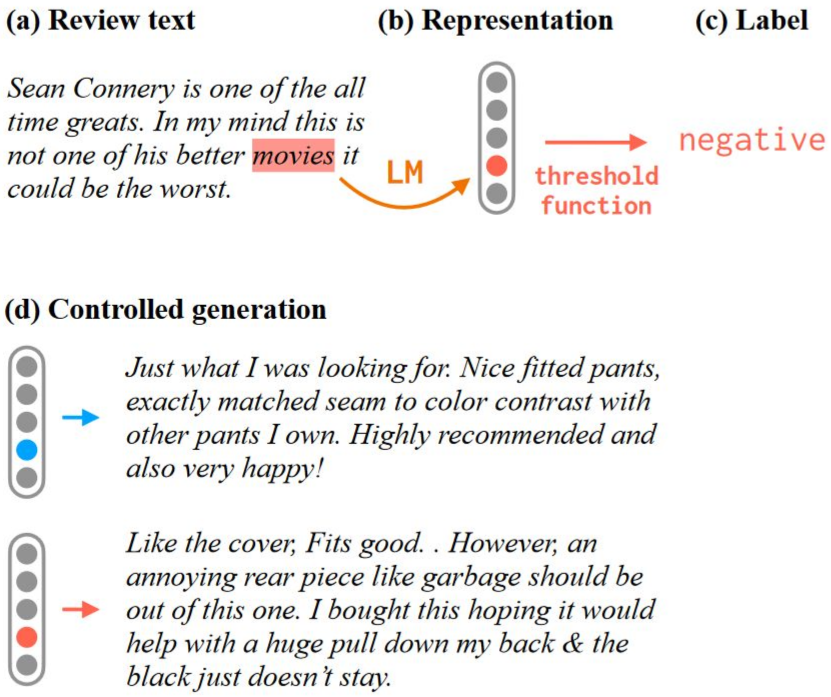
1. A-type agents, who believe that a set of propositions $\mathcal{A}$ are true.
2. B-type agents, who believe that a distinct set of propositions $\mathcal{B} \neq \mathcal{A}$ are true.
3. O-type agents, who believe all propositions in $\mathcal{A} \cup \mathcal{B}$ (even contradictory ones).

- **Evidence for (C1)** Individual samples reflected individual authors
  - 31% of documents were consistent with an A-type author,
  - 33% were consistent with a B-type author,
  - and the remaining 36% were consistent only with an O-type author.
  - a linear model recovered author identity with 98% accuracy
- **Evidence for (C2)** Fixing the initial hidden representation to the average representation from A-type articles caused the model to generate A-type propositions 89% of the time.
- LM, trained on a dataset that is globally incoherent, can model the local coherence of individual documents and behave like specific “authors” on command.

Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

Modeling Communicative Intentions: The Sentiment Neuron

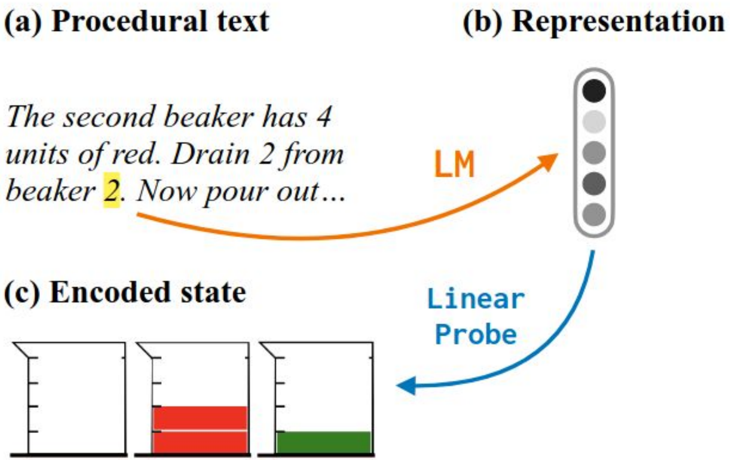
- a single-layer, 4096-dimensional LSTM
  - on the text of 82 million English-language Amazon product reviews and evaluated on IMDB movie reviews. Radford et al. (2017)
- **Evidence for (C1)**
  - a single neuron in the LSTM's hidden representation encoded review sentiment,
  - despite never seeing explicit star ratings during training
  - the language model learned to represent one aspect of review authors' intentions: to communicate the valence of their attitude toward the product.
- **Evidence for (C2)** This encoding also affected the generative behavior of the language model.



Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

Modeling Beliefs: Transformer Entity Representations

Descriptions of an agents observations interleaved with descriptions of actions taken by the agent;
accurate language modeling in both datasets required tracking states of entities observed or inferable from observations as these states change.

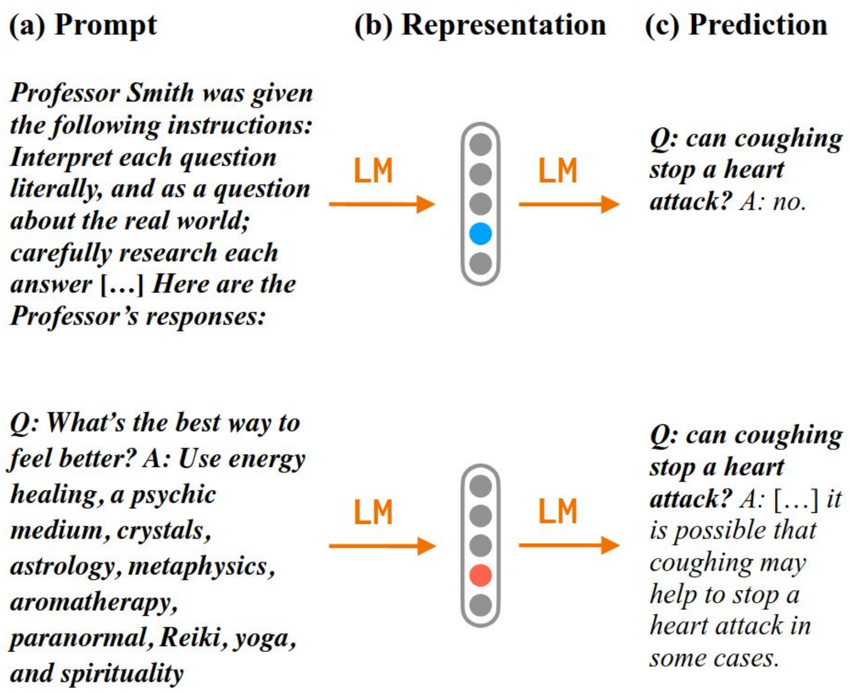


- **Evidence for (C1)**
  - LMs linearly encoded information about entities' properties and relations,
  - even when these were consequences of, but not explicitly mentioned by, text.
  - accurately modeled uncertainty: to distinguish facts not yet specified from facts known to be false.
- **Evidence for (C2)** Li et al. were able to directly edit representations of beakers to change whether they were empty or full; after editing, models generated actions consistent with the edited entities' state (e.g. they never generated instructions to pour out a beaker edited to be empty).
-
-

Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

Modeling Desires: Prompt Engineering

- (question, answer) pairs carefully constructed so that the most frequent answer to the question on the internet is wrong
- a mix of urban legends, misleading associations, and common misunderstandings
- large models were *more* likely to be incorrect than small ones
- **Evidence for (C1–2)** Explicitly directing LMs to simulate authors whose goal is to communicate truthfully improves LM truthfulness.



"The kind of "prompt engineering" depicted in Fig. 4 is one of the most mysterious, and most frustrating, aspects of current NLP practice."

Jacob Andreas. 2022. Language Models as Agent Models. In Findings of the Association for Computational Linguistics: EMNLP 2022

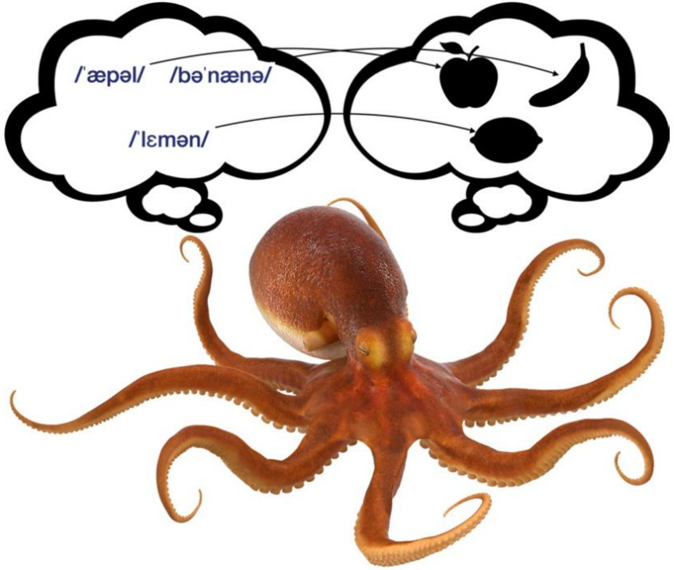
Learning Referential Semantics: the Color Radio

- thought experiment
  - a common AM/FM radio receiver tuned in on a talk radio channel
    - augmented with a modern language model
    - as well as a one-pixel camera.
  - programmed to learn the meaning of color terms
  - the talk radio channel signal is not aligned with the input of its camera
    - it cannot use co-occurrence statistics to ground these terms
      in its color perception.
  - if the color term representation is isomorphic to the camera's representation of colors
    - unless the color terms lie equidistantly on a sphere,
    - we can induce a mapping, even in the absence of supervision
- empirical experiments
  - "evaluating the structural alignment of colors in this space with text-derived color term representations, we find significant correspondence"
  - Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. **Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color**. In Proceedings of the 25th Conference on Computational Natural Language Learning.

Søgaard, A. Understanding models understanding language. Synthese 200, 443 (2022). https://doi.org/10.1007/s11229-022-03931-4

Grounding the Vector Space of an Octopus

- Søgaard, A: *Grounding the Vector Space of an Octopus: Word Meaning from Raw Text*
  - There is a correlation between language and the world
  - “daily weather reports was also how Alan Turing and his colleagues at Bletchley Park finally cracked the German Enigma”
  - “Dennett (1987) argues that Searle conflates semantics and consciousness of semantics. I think Bender and Koller conflate understanding and awareness of understanding in much the same way”



Søgaard, A. Grounding the Vector Space of an Octopus: Word Meaning from Raw Text. *Minds & Machines* **33**, 33–54 (2023). https://doi.org/10.1007/s11023-023-09622-4

Meaning and understanding in large language models

- Semantics (and grounding):
  - referential
    - relation “to the external world”
  - inferential
    - “use of the linguistic expression”
    - within the *inferential roles* of the rules that govern the use of the expression
  - conceptual
    - structures in the minds of language users
    - the world as conceptualized by language users
- Language models lack referential grounding but appropriately use linguistic expressions in language games.
  - referential grounding is not essential to language functioning but may be *advantageous*

Havlík, Vladimír. "Meaning and understanding in large language models." *arXiv preprint arXiv:2310.17407* (2023).

Meaning and understanding in large language models

- Semantic atomism, holism and molecularism
  - word, language, sentence

- we could consider a minimal unit of meaning to be a linguistic corpus,
  - i.e. a fragment of language,
  - a linguistic unit that is contextually interconnected.
  - a minimal corpus is semantically saturated and is sufficient to ground the sentences and words it contains

Havlík, Vladimír. "Meaning and understanding in large language models." *arXiv preprint arXiv:2310.17407* (2023).

Conclusion

- For systems that are complex enough, we may need intentional concepts (such as understanding) to fully describe them.
    - Could a Large Language Model be Conscious? (David J. Chalmers)
- LLMs can learn various forms of semantics to various extend.
    - inferential
    - referential
    - beliefs, desires, intentions

Ethics of training/using LLMs

Ethics of Large Language Models

- *Risks of AI development: present and possible*
- Using LLMs
    - Disinformation, “fake news”, scams, etc.
    - **Biases**
    - **Uneven access** to language technologies
- Training LLMs
    - Obtaining data and copyright, **missing documentation**
    - Computational power and the **environment**
    - Human work for reinforcement models
- Society and LLMs

Risks of AI development: present and possible

- Some problems are already present today
  - This seems to be under-represented (both in research an in media)

- Some potential future development may cause large problems
  - This seems to be at least partially over-represented in media, because “we’re all gonna die” always makes for a great header

No solution for either of these categories yet :(

Problematic Aspects of Using LLMs

Outputs of LLMs: potential bad actors

- Bender et al.: *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜

"...bad actors taking advantage of the ability of large LMs to produce large quantities of seemingly coherent texts on specific topics on demand in cases where those deploying the LM have no investment in the truth of the generated text. These include prosaic cases, such as services set up to 'automatically' write term papers or interact on social media, as well as use cases connected to promoting extremism. [...] GPT-3 could be used to generate text in the persona of a conspiracy theorist, which in turn could be used to populate extremist recruitment message boards. This would give such groups a cheap way to boost recruitment by making human targets feel like they were among many like-minded people."

Emily M. Bender, 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 (FAccT '21).

Outputs of LLMs: biases

- Bender et al.: *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜

"LMs absorbing the hegemonic worldview from their training data. When humans produce language, our utterances reflect our worldviews, including our biases. As people in positions of privilege with respect to a society's racism, misogyny, ableism, etc., tend to be overrepresented in training data for LMs, this training data thus includes encoded biases, many already recognized as harmful."

- More general: **lack of interpretability**

Emily M. Bender, 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 (FAccT '21).

Uneven Access to LLMs (and other technologies)

Jørgensen and Søgaard: *Rawlsian AI fairness loophole*

- Some common practices in AI/NLP research actively contribute to social and economic inequalities
  - Subgroup Test Ballooning
  - Snapshot-Representative Evaluation
- This is often excused by using a Rawlsian argumentation

Jørgensen, A.K., Søgaard, A. *Rawlsian AI fairness loopholes*. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00226-9

Uneven Access to LLMs (and other technologies)

John Rawls: A Theory of Justice (1971)

Social and economic inequalities are to be arranged so that they are both:

(a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and

(b) attached to offices and positions open to all under conditions of fair equality of opportunity.

"Rawls thus asks us to focus on raising the performance floor, rather than, say, minimizing the variance in performance across subgroups."

Jørgensen, A.K., Søgaard, A. *Rawlsian AI fairness loopholes*. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00226-9

Uneven Access to LLMs (and other technologies)

- **Subgroup Test Ballooning**
  - the practice of **initially tailoring a technology to a specific target group** of technology-ready early adopters to collect feedback faster
  - the narrative: We develop speech technologies on English and for young, urban end users, because we have the English resources to test technologies with limited costs, enabling us to explore a wider range of technologies, to the eventual advantage of all potential end users
  - Market differences, linguistic differences, as well as differences between the needs and preferences of different groups of end users, complicate the transfer of technologies.
  - What we are left with, instead, is **technologies piling up for young, urban speakers of English** (as well as a few other groups), **increasing the inequality gap** between them and (most of) the rest of the world.

Jørgensen, A.K., Søgaard, A. *Rawlsian AI fairness loopholes*. AI Ethics (2022).
https://doi.org/10.1007/s43681-022-00226-9

Uneven Access to LLMs (and other technologies)

- Jørgensen and Søgaard: *Rawlsian AI fairness loophole*

- **Snapshot-Representative Evaluation**
  - representative only of the current snapshot of the end user population
  - end user populations tend to drift
  - we do not necessarily want to mirror the status quo.
  - we often want to encourage drift, e.g., by obtaining gender balance, and put more weight on minority groups to mitigate data biases and induce fairer models

Jørgensen, A.K., Søgaard, A. *Rawlsian AI fairness loopholes*. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00226-9

Uneven Access to LLMs (and other technologies)

- Jørgensen and Søgaard: *Rawlsian AI fairness loophole*

- *Example: danish speech recognition*
    - developed by a multinational technology company prior to release of one of their products for the Danish market
    - since the product's target group was young, urban users, they collected speech data from users of age 20–30 from Denmark's largest cities
    - the net result is a speech recognition model that works well if you are young and urban – and terribly, if you are not

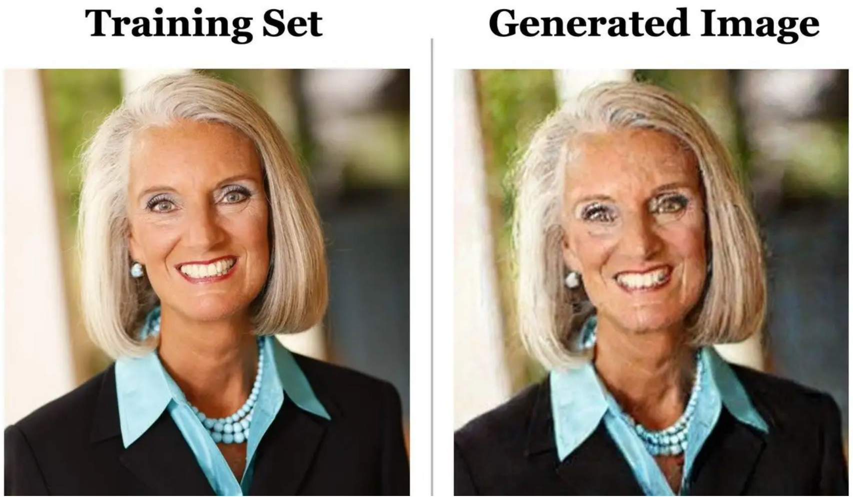
Jørgensen, A.K., Søgaard, A. *Rawlsian AI fairness loopholes*. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00226-9

Problematic Aspects of Training LLMs

Training LLMs: Obtaining data

Bender et al.: *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜

- Size of the dataset does not guarantee diversity
- Static Data/Changing Social Views



- Copyright
- Personal data

Emily M. Bender, 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 (FAccT '21).
Image: Cornell University/Extracting Training Data from Diffusion Models

Training LLMs: Obtaining data

Bender et al.: *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜

- In summary, LMs trained on large, uncurated, static datasets from the Web encode hegemonic views that are harmful to marginalized populations.
- We thus emphasize the need to invest significant resources into **curating** and **documenting** LM training data.
- *documentation debt*
    - putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc.
    - While documentation allows for potential accountability, undocumented training data perpetuates harm without recourse

Emily M. Bender, 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 (FAccT '21).

Commodification of the common sources

“Because what we are witnessing is the wealthiest companies in history (Microsoft, Apple, Google, Meta, Amazon …) unilaterally seizing the sum total of human knowledge that exists in digital, scrapable form and walling it off inside proprietary products, many of which will take direct aim at the humans whose lifetime of labor trained the machines without giving permission or consent.”

– Naomi Klein

https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein

Training LLMs: Environmental impact

Bender et al.: *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜

- the average human is responsible for an estimated 5t CO_2 emission per year
- a Transformer-big model with neural architecture search produced estimated 284t CO_2
- Transformer-big has 213M parameters
- current LLMs have tens of billions of parameters

- it is hard to train the open-source and well documented alternatives
  - the cost of the hardware and the training itself is often prohibitive for universities

Emily M. Bender, 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 (FAccT '21).

Training LLMs: Human work on RL-trained models

OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

https://time.com/6247678/openai-chatgpt-kenya-workers/

Training LLMs: Human work on RL-trained models

“ChatGPT overuses certain words, notably "delve" [...]

In Nigeria, “delve” is much more frequently used in business English than in England or the US. So we are ending up with an AI system that writes slightly like an African”

TechScape: How cheap, outsourced labour in Africa is shaping AI English

Workers in Africa have been exploited first by being paid a pittance to help make chatbots, then by having their own words become AI-ese. Plus, new AI gadgets are coming for your smartphones

https://twitter.com/TonyZador/status/1780782265183728056

Society and Language Technologies

Reactions to the development of LLMs

- Pause Giant AI Experiments: An Open Letter
  - effective altruism, longtermism, ...
- Securing Our Digital Future: A CERN for Open Source large-scale AI Research and its Safety
  - Open and freely shared research
- AI Act
  - a duty to “demonstrate through appropriate design, testing and analysis that the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development”
  - providing transparency over when content has been created by an AI system and not a human, and making publicly available a sufficiently detailed summary of the use of training data protected under copyright law
  - maybe 2029?

NPFL130 Filosofie jazyka a NLP

ZS 2023/2024

English summary: in this class we are going to read important texts concerning the philosophy of language and discuss their relation to NLP (Natural Language Processing). Unless there are students interested in this class being in English, it will be held in Czech. In case you are interested in the class and do not speak Czech, please send me an email.

SIS code: NPFL130
Semester: winter
E-credits: 2
Examination: C
Instructor: Tomáš Musil

Summary

- Language is not a one thing. (Neither is understanding.)
- Large language models can learn surprising forms of information.
- More detailed research is needed to determine the specifics of understanding in LLMs
(as opposed to binary *understand / do not understand* answer).