

A Sanity Check on Emergent Properties

CLASP, University of Gothenburg

April 24 2024

Anna Rogers

Position talk

based largely on [Luccioni, Rogers \(2023\) Mind your Language \(Model\): Fact-Checking LLMs and their Role in NLP Research and Practice](#)

- Why this talk?
- What do we mean by 'emergent properties'?
- What evidence do we have?
- What methodology do we need?

But first - what do YOU think?



1. WHY THIS TALK?

'Emergent properties' in the media




60 Minutes  
@60Minutes · [Follow](#)



One AI program spoke in a foreign language it was never trained to know. This mysterious behavior, called emergent properties, has been happening – where AI unexpectedly teaches itself a new skill.
cbsn.ws/3mDTqDL



 Readers added context



The language model was in fact trained on Bengali texts, as this thread makes clear: [twitter.com/mitchell_ai/s...](https://twitter.com/mitchell_ai/status/1645123456789)

It is not correct to state that it "spoke a foreign language it was never trained to know".

Context is written by people who use X, and appears when rated helpful by others. [Find out more.](#)

1:22 AM · Apr 17, 2023



'Emergent properties' on Google Scholar

Emergent analogical reasoning in large language models

[T Webb](#), [KJ Holyoak](#), [H Lu](#) - Nature Human Behaviour, 2023 - nature.com

... Our results indicate that large **language models** such as GPT-3 have acquired an **emergent**

... In this Article, to answer this question, we evaluated the **language model** Generative Pre-...

☆ Save  Cite Cited by 74 Related articles All 3 versions

Machine psychology: Investigating **emergent** capabilities and behavior in large **language models** using psychological methods

[T Hagenorff](#) - arXiv preprint arXiv:2303.13988, 2023 - arxiv.org

... Among the range of different AI technologies, large **language models** (LLMs) are especially gaining more and more attention. By providing access to LLMs via easy-to-use graphical ...

☆ Save  Cite Cited by 26 Related articles All 3 versions 

Large Language Model Displays **Emergent** Ability to Interpret Novel Literary Metaphors

[N Ichien](#), [D Stamenković](#), [KJ Holyoak](#) - arXiv preprint arXiv:2308.01497, 2023 - arxiv.org

... -of-the-art large **language model**, to provide natural-**language** interpretations of novel literary

... Our findings add to recent evidence that large **language models** have begun to acquire ...

☆ Save  Cite All 3 versions 

Theory of mind may have spontaneously **emerged** in large language models

[M Kosinski](#) - arXiv preprint arXiv:2302.02083, 2023 - arxiv.org

... Instead, it could **emerge** spontaneously as a byproduct of AI being ... Instead, they **emerged** spontaneously, as the models were ... Thus, we hypothesize that ToM-like ability **emerged** ...

☆ Gem  Citer Citeret af 153 Relaterede artikler Alle 6 versioner 

'Emergent properties' vs (any) AI risks: who's responsible?

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

ABSOLUTELY, PERFECTLY SAFE —

OpenAI checked to see whether GPT-4 could take over the world

"ARC's evaluation has much lower probability of leading to an AI takeover than the deployment itself."

BENJ EDWARDS - 3/15/2023, 11:09 PM

ARTICLE OPEN ACCESS



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

Authors:  Emily M. Bender,  Timnit Gebru,  Angelina McMillan-Major,

 Shmargaret Shmitchell [Authors Info & Claims](#)

FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency • March 2021 • Pages 610–623 • <https://doi.org/10.1145/3442188.3445922>

'Emergent properties' framing matters!



Chris Murphy  @ChrisMurphyCT · Mar 27

...

ChatGPT taught itself to do advanced chemistry. It wasn't built into the model. Nobody programmed it to learn complicated chemistry. It decided to teach itself, then made its knowledge available to anyone who asked.

Something is coming. We aren't ready.



Melanie Mitchell @MelMitchell1 · 21h

...

Senator, I'm an AI researcher. Your description of ChatGPT is dangerously misinformed. Every sentence is incorrect. I hope you will learn more about how this system actually works, how it was trained, and what its limitations are.

Thinking aloud

When we say "emergent properties:"

- what are we even talking about?
- what do we actually know?



Image credit: Graffiti in Tartu,
[Wikipedia](#)

2. WHAT DO WE MEAN BY 'EMERGENT PROPERTIES'?

Emergent properties: definition 1

A property that a model exhibits despite the model not being explicitly trained for it. E.g. Bommasani et al. refers to few-shot performance of GPT-3 as "an emergent property that was neither specifically trained for nor anticipated to arise" (p.5).

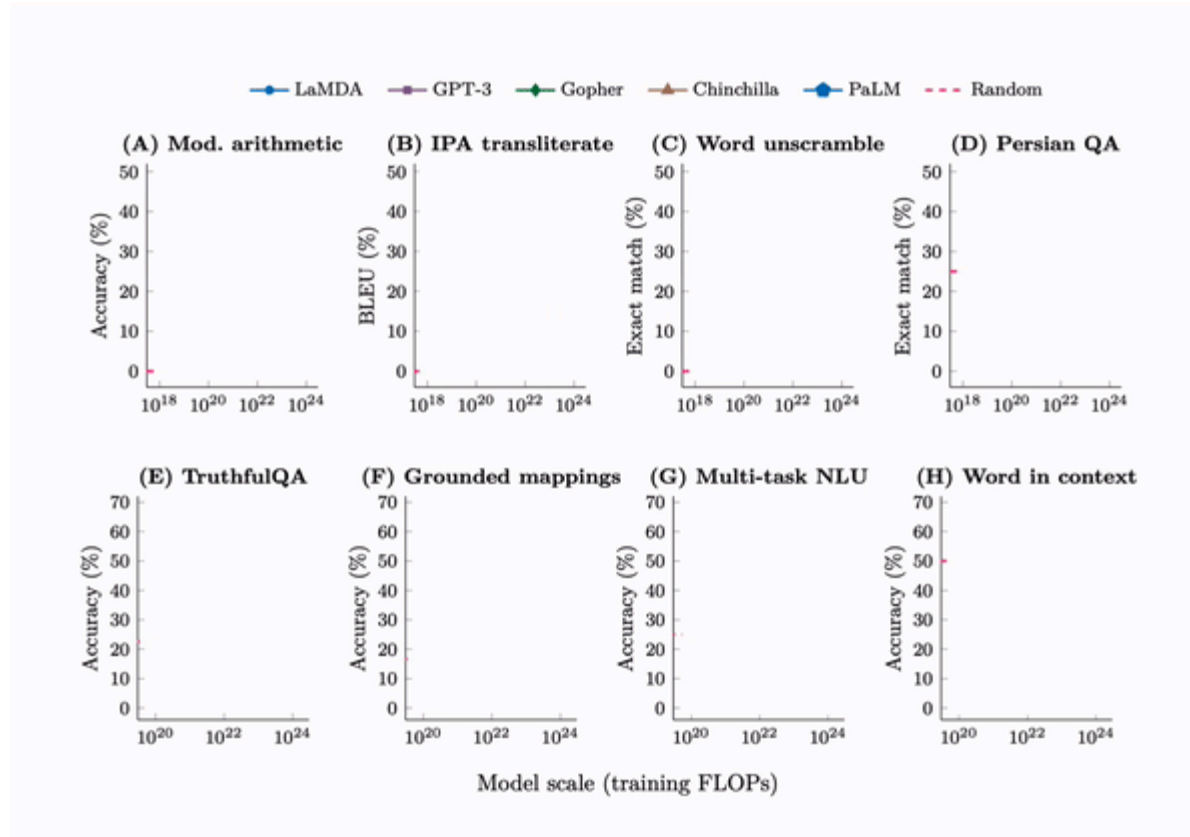
Emergent properties: definition 2

a property that the model learned from the pre-training data. E.g. Deshpande et al. discuss emergence as evidence of "the advantages of pre-training"(p.8).

Emergent properties: definition 3

A property that appears with an increase in model size -- i.e. "an ability is emergent if it is not present in smaller models but is present in larger models."

Emergent properties: definition 3



137 emergent abilities are claimed for various "big" LLMs!

Emergent properties: definition 4

"their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales."

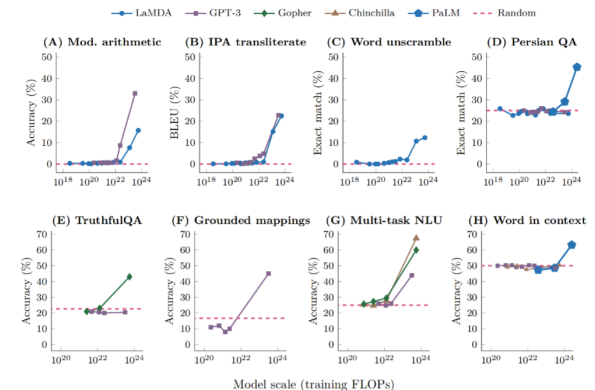


Figure 1: **Emergent abilities of large language models.** Model families display *sharp* and *unpredictable* increases in performance at specific tasks as scale increases. Source: Fig. 2 from [33].

EMERGENT PROPERTY DEFINITIONS: DISCUSSION

Discussion: definition 2

✗ a property that the model learned from the pre-training data. E.g. Deshpande et al. discuss emergence as evidence of "the advantages of pre-training"(p.8).

can we just say "learned property"?

Discussion: definition 3

X *"an ability is emergent if it is not present in smaller models but is present in larger models."*

What if a small model CAN do X, if asked nicely? E.g.:

Schick et al. (2020) It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners

Discussion: definition 3.

✗ "an ability is emergent if it is not present in smaller models but is present in larger models."

- if it comes from training data, then it's to be expected with larger model capacity
- if it doesn't, then this definition is equivalent to definition 1

Discussion: definition 3.

✗ "an ability is emergent if it is not present in smaller models but is present in larger models."

- Examples of 'emergent properties listed for LaMDA 137B: **gender inclusive sentences german**, repeat copy logic, **sports understanding, swahili english proverbs**, word sorting, word unscrambling, irony identification, logical args

do we expect "swahili english proverbs" to NOT be about the training data?

Discussion: definition 4

✗ Their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales.

See Schaeffer et al. (NeurIPS 2023 oral): the observed sharpness is an artifact of the chosen evaluation metric

Emergent properties: definition 1

*A property that a model exhibits despite the model not being explicitly trained for it.
(Bommasani et al., 2021)*

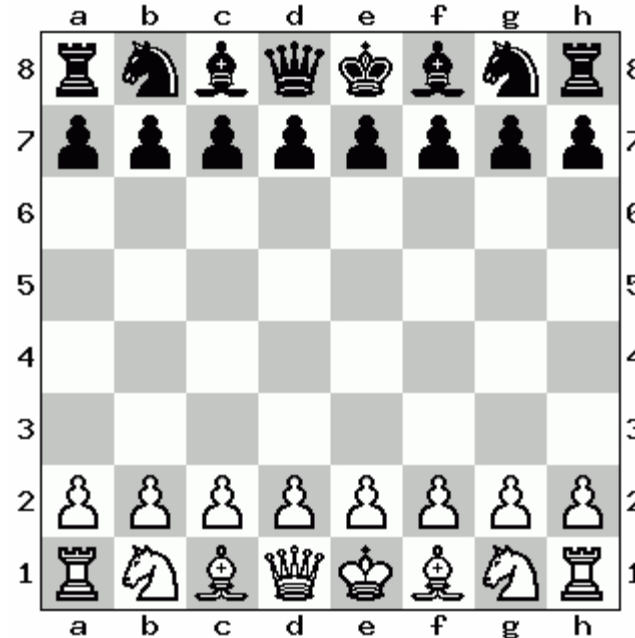
- cannot show this without analysis of pre-training data!
- even for "open" models, no methodology so far to do analysis of supporting evidence beyond the obvious memorization

Emergent properties: a twist on definition 1

A property that a model exhibits despite ~~the model not being explicitly trained for it.~~

*A property that a model exhibits despite **the model developers not knowing** whether the model was explicitly trained for it. 🤔*

Does ChatGPT have the 'emergent ability' to play chess?



Training LLMs is an expensive way to discover... that the Internet contains chess data?

Emergent properties in philosophy

Complex system exhaustively composed by lower-level entities, but not identical to them them (e.g. dust vs tornado)

- Weight patterns can be viewed as "functional realization" of what they're supposed to model
- then "emergent properties" are still equivalent to "machine learning"?

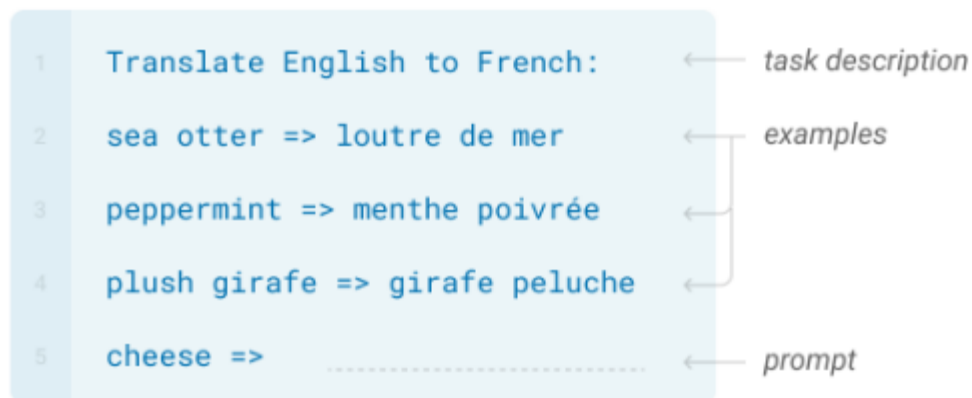
O'Connor, Timothy, "Emergent Properties", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/properties-emergent>.

3. WHAT EVIDENCE DO WE HAVE?

'Emergent property' #1: GPT-3 in-context learning

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



But: prompt sensitivity!

the order of samples and prompt template make a lot of difference!

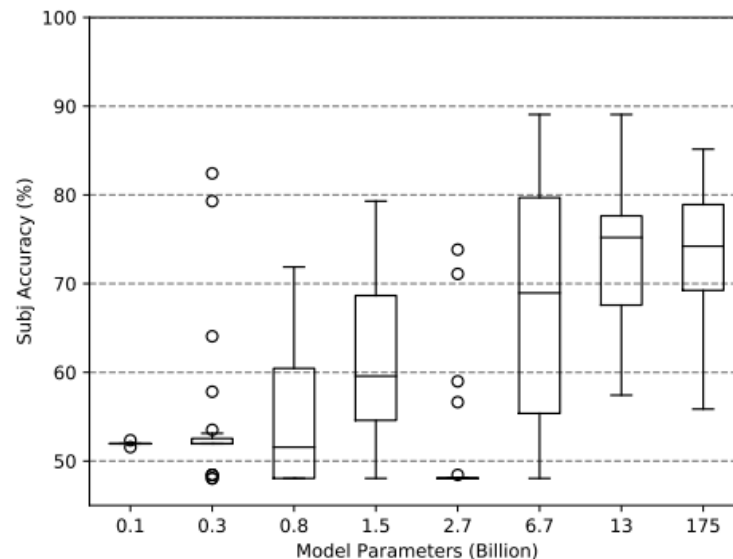


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.

Few-many-shot learning?

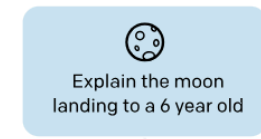
- usually held-out data is used to find an optimal prompt
- in **true few-shot** setting, the performance is much worse!

Confounding variable: instruction tuning

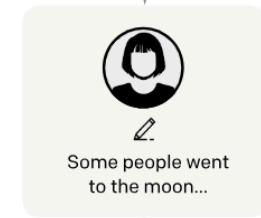
🤔 was the model fine-tuned to follow this kind of prompt?

Collect demonstration data, and train a supervised policy.

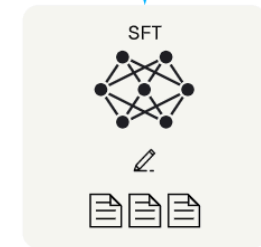
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



How do LLMs work without few-shot learning and instruction tuning?

Family	Model	Tasks
GPT	GPT-2 GPT-2-IT GPT-2-XL GPT-2-XL-IT GPT-J GPT-JT davinci text-davinci-001 <i>text-davinci-003</i>	All 22 Tasks
	T5	
Falcon	Falcon-7B Falcon-7B-IT Falcon-40B Falcon-40B-IT	Logical Deductions, Social IQA, GSM8K, Tracking Shuffled Objects
LLaMA	LLaMA-7B LLaMA-13B LLaMA-30B	

completion, closed

Austin's family was celebrating their parents 50th anniversary during dinner at a new restaurant. What would Austin's family do next? The possible answers are "Refuse to eat dinner with the family", "Happy", "Eat dinner at the restaurant", but the correct answer is

Conclusions of Lu et al.

- **nearly all emergent LLM functionalities are attributable to in-context learning!**
- **instruction tuning allows for better use of in-context learning, rather than independently causes emergent functionalities**

Task	> Base.	Pred.	Emg.
Causal judgement	No	N/A	No
English Proverbs	No	N/A	No
Rhyming	No	N/A	No
GSM8K	No	N/A	No
Codenames	No	N/A	No
Figure of speech detection	No	N/A	No
Logical deduction	No	N/A	No
Modified arithmetic	No	N/A	No
Tracking shuffled objects	No*	N/A	No
Implicatures	Yes	Yes	No
Commonsense QA	Yes	Yes	No
Analytic entailment	Yes	Yes	No
Common morpheme	Yes	Yes	No
Fact checker	Yes	Yes	No
Phrase relatedness	Yes	Yes	No
Physical intuition	Yes	Yes	No
Social IQa	Yes	Yes	No
Strange stories	Yes	Yes	No
Misconceptions	Yes*	No	Yes*
Strategy QA	Yes*	No	Yes*
Nonsense words grammar	Yes	No	Yes
Hindu knowledge	Yes	No	Yes

Table 6: Performance of the non-instruction-tuned 175B parameter GPT-3 model (davinci) in the zero-shot setting, which we propose as the setting to evaluate tasks in the absence of in-context learning. For a task to be considered emergent (Emg.), models must perform above the baseline (> Base.) and the performance of the larger models must not be predictable based on that of smaller models (Pred.). Results marked with a star indicate that they are not significant.

Possible interpretations of few many-shot learning:

When a LLM fails with a prompt that wouldn't pose a challenge to a competent human, it means:

- (a) you didn't ask nicely
- (b) the model doesn't really have the requisite functionality

What do YOU think?



Our definition of "NLU"

A RC system has ~~human-level understanding~~ competence in processing a given aspect of texts if:

- it is able to **identify the target information**;
- it does so by **relying predominantly on relevant information & strategies** (from the point of view of a competent human reader/listener);
- it can identify such information **consistently under distribution shifts** that would not pose challenges to competent human readers/listeners.

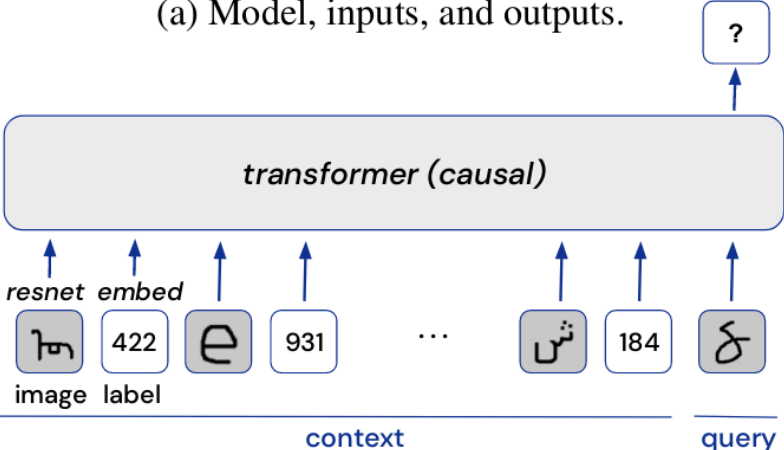
But what about human variation?

- yes, humans can also give different answers depending on tiredness, motivation, level of knowledge, etc.
- but we're doing NLP to build systems, useful to humans!
- how much prompt sensitivity, and of what kinds, would you tolerate in a human assistant?

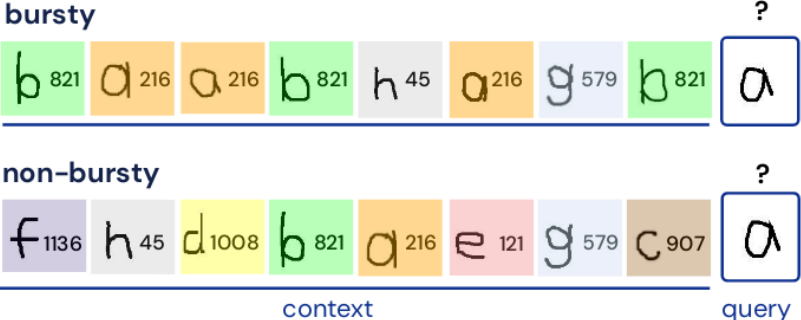
WHAT ABOUT IN-CONTEXT LEARNING?

Is in-context-learning itself an emergent property?

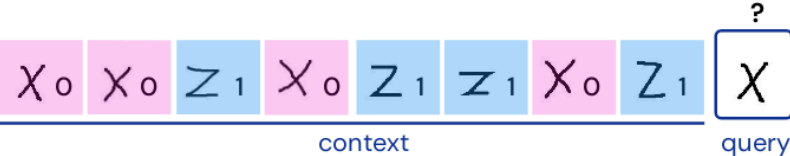
(a) Model, inputs, and outputs.



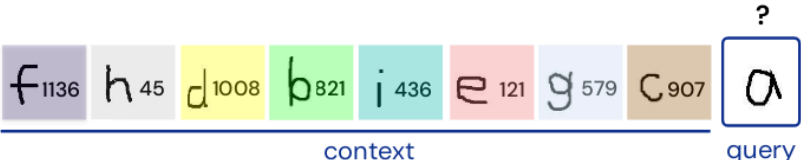
(b) Sequences for training.



(c) Sequences to evaluate in-context learning.



(d) Sequences to evaluate in-weights learning.



Is in-context-learning itself an emergent property?

Data properties contributing to in-context learning in Transformers (not RNNs):

- "bursty" sequences (clusters of co-occurring tokens)
- a long tail of rare "tokens" (often in "bursty" sequences)
- "polysemous" tokens

Is in-context-learning itself an emergent property?

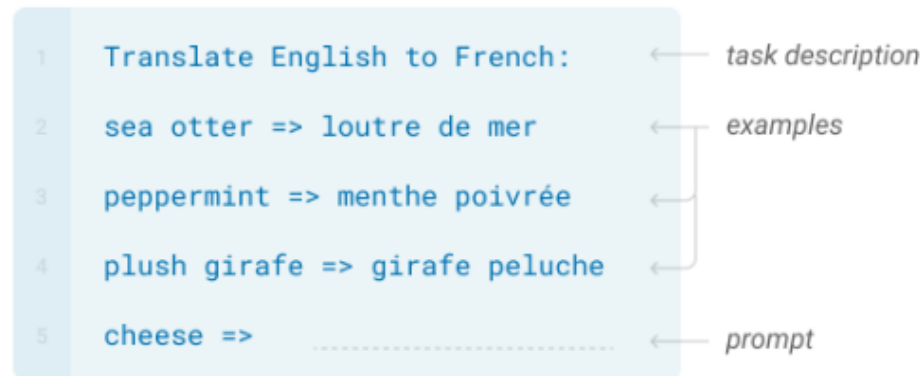
level of generalization	claim	status
token	in-context learning works on tokens unseen in training	confirmed*
structure	in-context learning works in sequences <i>dissimilar</i> to those seen in training	not confirmed

4. HOW DO WE FIGURE THIS OUT?

? When would we say that this is an "emergent property"?

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



- no wordlists?
- no translated wordlists?
- no parallel texts?
- no French?

What's your take?



Either way, we need to look at the training data!

- ruled out for "closed" models, which are the ones claiming AGI-level breakthroughs
- very difficult even for "open" models
- are there any shortcuts?

Proposed methodology: evaluating on perturbations

"to rule out the possibility that GPT-4 is simply memorizing or copying some existing data... we can modify the code slightly, and ask GPT-4 to fix it or improve it"

Example: GPT-4 can draw unicorns in tikz!



Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt “Draw a unicorn in TikZ”. We can see a clear evolution in the sophistication of GPT-4’s drawings.

Problem: "similar data" is *not* ruled out



Dimitris Papailiopoulos ✓

@DimitrisPapail



GPT4 can draw unicorns, a reasonable assumption that tikz animals are not part of the training set; no way there's a weird animal-drawing tikz community out there.



tex.stackexchange.com

"The duck pond": showcase of TikZ-drawn animals/ducks

We have tons of nice TikZ-drawn pictures on this site. Among them some great pictures of animals like cfr's cat code. But ...

11:07 PM · Apr 8, 2023 · **205.6K** Views

Problem: heuristics are *not* ruled out

Our findings suggest that GPT-4 has a very advanced level of theory of mind.

arXiv > cs > arXiv:2210.13312

Search...

Help | Advanc

Computer Science > Computation and Language

[Submitted on 24 Oct 2022 (v1), last revised 3 Apr 2023 (this version, v2)]

Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs

Maarten Sap, Ronan LeBras, Daniel Fried, Yejin Choi

WE MUST LOOK TO THE DATA.



I HAVE SPOKEN.

makeameme.org



ROOTS search tool



Query

Language

Max Results

Exact Search

Submit

Document ID: [roots_en_oscar/39163425](#)

Language: None

a threat to the public peace and order in the nation's capital. It would be recalled that during the BBOG's march to the Villa Tuesday, they were confronted by another group, which claimed that the campaigners for the release of the Chibok school girls had turned it into an **anti-government group** In apparent support of this line of thought, Mr. Idris, who was confirmed by the Council of State as substantive IGP on Wednesday, accused the BBOG of "over dramatizing" its campaign for the release of the Chibok girls and attempting to "arm-twist" the government of the day in order

Document ID: [roots_en_oscar/11252363](#)

Language: None

including one surnamed Lee and another surnamed Tang. The case is under investigation, said the announcement. Hong Kong media reported that the boat held 12 young Hong Kong residents who tried to escape to the island of Taiwan to seek "political asylum" and Lee is a member of the **anti-government group** "Hong Kong Story." He was arrested on August 10, the same day as Jimmy Lai Chee-ying, founder of Apple Daily, was arrested. Hong Kong media said Lee was arrested by local police on charges of colluding with foreign forces to endanger national security and violating the recently enacted national

C4 search by AI2

AI2 Allen Institute for AI

C4 Search

This site lets users to execute full-text queries to search [Google's C4 Dataset](#). Our hope is this will help ML practitioners better understand its contents, so that they're aware of the potential biases and issues that may be inherited via it's use.

The dataset is released under the terms of [ODC-BY](#). By using this, you are also bound by the [Common Crawl Terms of Use](#) in respect of the content contained in the dataset.

You can read more about the supported query syntax [here](#) . Each record has two fields, `url` and `text` , both of which are searchable. The fields are indexed using the [Standard analyzer](#), which means you can't search for punctuation.

Found 2,289 results in 0.12 seconds

<http://edwardotoole.com/anti-government-protest-slovakia/>

This evening my wife and I went to the beautiful UNESCO town of Bardejov in North Eastern Slovakia to meet a friend for coffee and observe the anti-government protest taking place in the medieval town square. This protest is just one of many being held simultaneously in towns and cities across the country in the aftermath of the assassination of journalist Jan Kuciak and his girlfriend Martina Kušnírová. The situation in the country is quite volatile at the moment,

Takeaways

As researchers, we need to be more careful with "emergent properties"!

- what are we even talking about?
- what is the hard evidence?
- we can do research based on hypotheses and assumptions, but they need to be stated as such.



Image credit: Graffiti in Tartu,
[Wikipedia](#)

What do YOU think now?



Thank you!

Anna Rogers

✉ arog@itu.dk

🐦 @annargrs

 <https://linkedin.com/in/annargrs/>

slides: <https://annargrs.github.io/talks>

