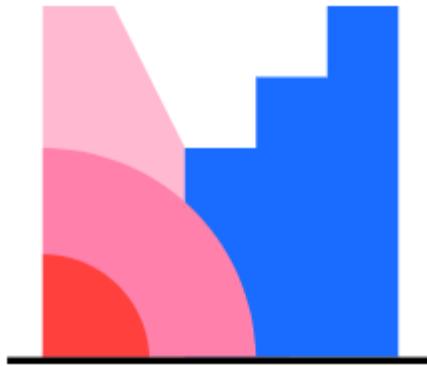


Part 2. Facts *on* LLMs

- What do benchmark numbers tell us?
- Can LLMs process long contexts well?
- Do LLMs have emergent properties?

WHAT DO BENCHMARK NUMBER TELL US?

Are you happy with current evaluations? Why/why not?



With 'closed' LLMs, we could be doing this:

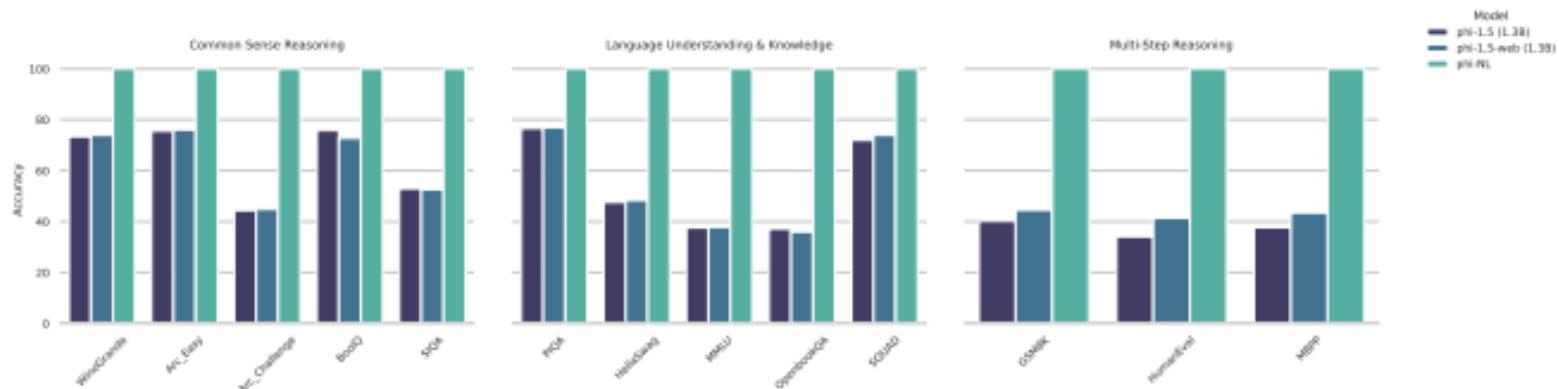
Pretraining on the Test Set Is All You Need

Rylan Schaeffer

September 19, 2023

Abstract

Inspired by recent work demonstrating the promise of smaller Transformer-based language models pretrained on carefully curated data, we supercharge such approaches by investing heavily in curating a novel, high quality, non-synthetic data mixture based solely on evaluation benchmarks. Using our novel dataset mixture consisting of less than 100 thousand tokens, we pretrain a 1 million parameter transformer-based LLM **phi-CTNL** (pronounced “fictional”) that achieves perfect results across diverse academic benchmarks, strictly outperforming all known foundation models. **phi-CTNL** also beats power-law scaling and exhibits a never-before-seen grokking-like ability to accurately predict downstream evaluation benchmarks’ canaries.



Schaeffer R. (2023) [Pretraining on the Test Set Is All You Need](#) (satire)

Caveat 1: test contamination!

ways of testing: dataset search, data extraction from models

Rows: 9

Corpus ↓	Dataset	Train split	Dev split	Test split	Source
The Pile	MNLI-m		2.2% Contaminated	N/A	Paper
The Pile	MNLI-mm		2.1% Contaminated	N/A	Paper
RedPajama	MNLI-m		2.3% Contaminated	N/A	Paper
RedPajama	MNLI-mm		2.2% Contaminated	N/A	Paper
OSCAR	MNLI-m		1.8% Contaminated	N/A	Paper
OSCAR	MNLI-mm		1.9% Contaminated	N/A	Paper
ChatGPT	MNLI-m	Contaminated	Contaminated	N/A	
C4	MNLI-m		1.6% Contaminated	N/A	Paper
C4	MNLI-mm		1.7% Contaminated	N/A	Paper

<https://hitz-zentroa.github.io/lm-contamination/>

Contamination does impact evaluation scores

EPG: 'estimated performance gain'

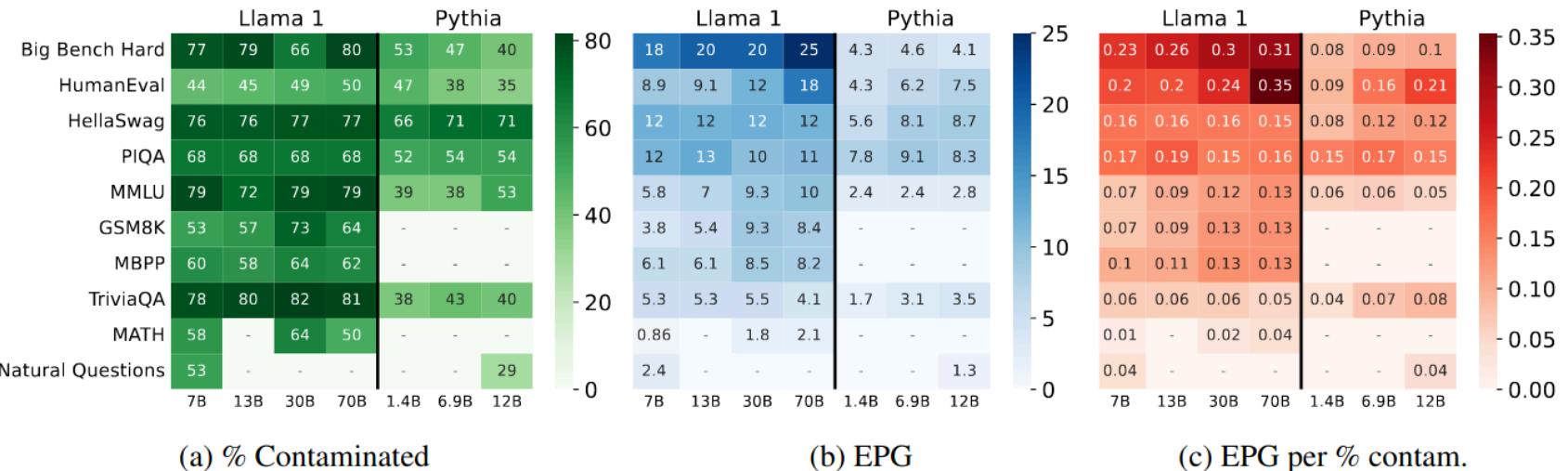


Figure 3: **Per model % contaminated and EPG values across benchmarks.** Percentage of the dataset marked contaminated (a) and corresponding EPG (b) and average gain per % contamination (c) for each of the model-benchmark pairs we considered in our study, according to the best contamination metric (see Table 3). Optimal thresholds are selected separately for each model-benchmark pair. For COPA, DM Contest, and SiQA no significant EPG was found, and they are therefore omitted from the plot.

Even worse: 'subtle' contamination

Answer	Test Question	Train Question
Jason Marsden	who plays max voice in a goofy movie	who does max voice in a goofy movie
January 23 2018	when will the 2018 oscar nominations be announced	when are the oscar nominations for 2018 announced
Alan Shearer	who has scored more goals in the premier league	most goals scored by a premier league player
retina	where are the cones in the eye located	where are cone cells located in the eye
francisco pizarro	who led the conquest of the incas in south america	conquistador who defeated the incan empire in peru

Dataset	% Answer overlap	% Question overlap
NaturalQuestions	63.6	32.5
TriviaQA	71.7	33.6
WebQuestions	57.9	27.5

Lewis et al. (2021) [Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets](#)

... now with a commercial COI?

... while OpenAI has access to FrontierMath, it has a “verbal agreement” with Epoch AI not to use FrontierMath’s problem set to train its AI.

Wiggers K. (2025) [AI benchmarking organization criticized for waiting to disclose funding from OpenAI](#)

? If we making new data, can we assume its originality?



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.

Bubeck et al. (2023) [Sparks of Artificial General Intelligence: Early experiments with GPT-4](#)

However...



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

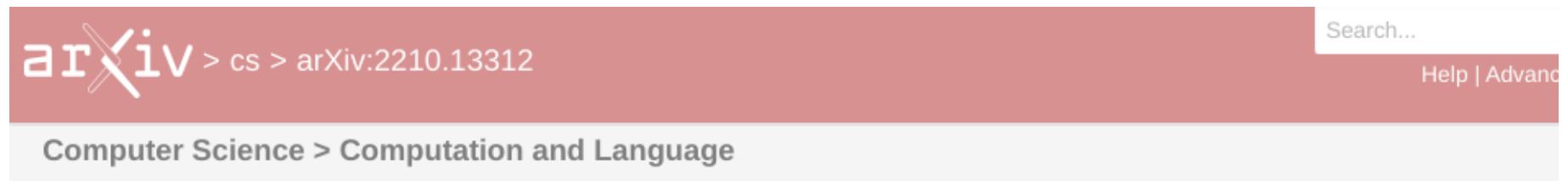
We have tons of nice TikZ-drawn pictures on this site. Among them some great pictures of animals like cfr's cat code. But ...

11:07 PM · Apr 8, 2023 · 205.6K Views

<https://twitter.com/DimitrisPapail/status/1644809234431848450?s=20>

Example 2: conclusions based on 'new' data

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



The screenshot shows the arXiv search interface. At the top, there is a red header bar with the arXiv logo and navigation links for 'Search...', 'Help | Advanced', and 'Logout'. Below the header, a search bar contains the query 'arXiv:2210.13312'. The main content area displays the results for this query, showing a single result from the 'Computer Science > Computation and Language' category. The title of the paper is 'Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs', and it was submitted on Oct 24, 2022, and last revised on Apr 3, 2023. The authors listed are Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi.

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](#)

Bubeck et al. (2023) [Sparks of Artificial General Intelligence: Early experiments with GPT-4](#)

Caveat2: fine-tuning vs few-shot performance

- generally, more in-domain training -> higher performance
- after GPT-3, most "big" models were presented with few-shot evaluations only
- hence, one could probably get higher performance on many tasks, than what is reported in most recent papers

e.g. superGLUE leaderboard: fine-tuned RoBERTa - 84.6,
few-shot GPT-3 - 71.8

<https://super.gluebenchmark.com/leaderboard/>

Caveat2: *True* few-shot performance

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

Perez et al. (2021) [True Few-Shot Learning with Language Models](#)

Caveat 3: are we measuring what we think we're measuring?

fine-tuning LMs on a range of NLP tasks, with instructions, improves their downstream performance on held-out tasks, both in the zero-shot and few-shot settings



among held-out tasks:

- follow instructions in 'non-English' languages
- perform summarization and question-answering for code

Caveat 3: are we measuring what we think we're measuring?

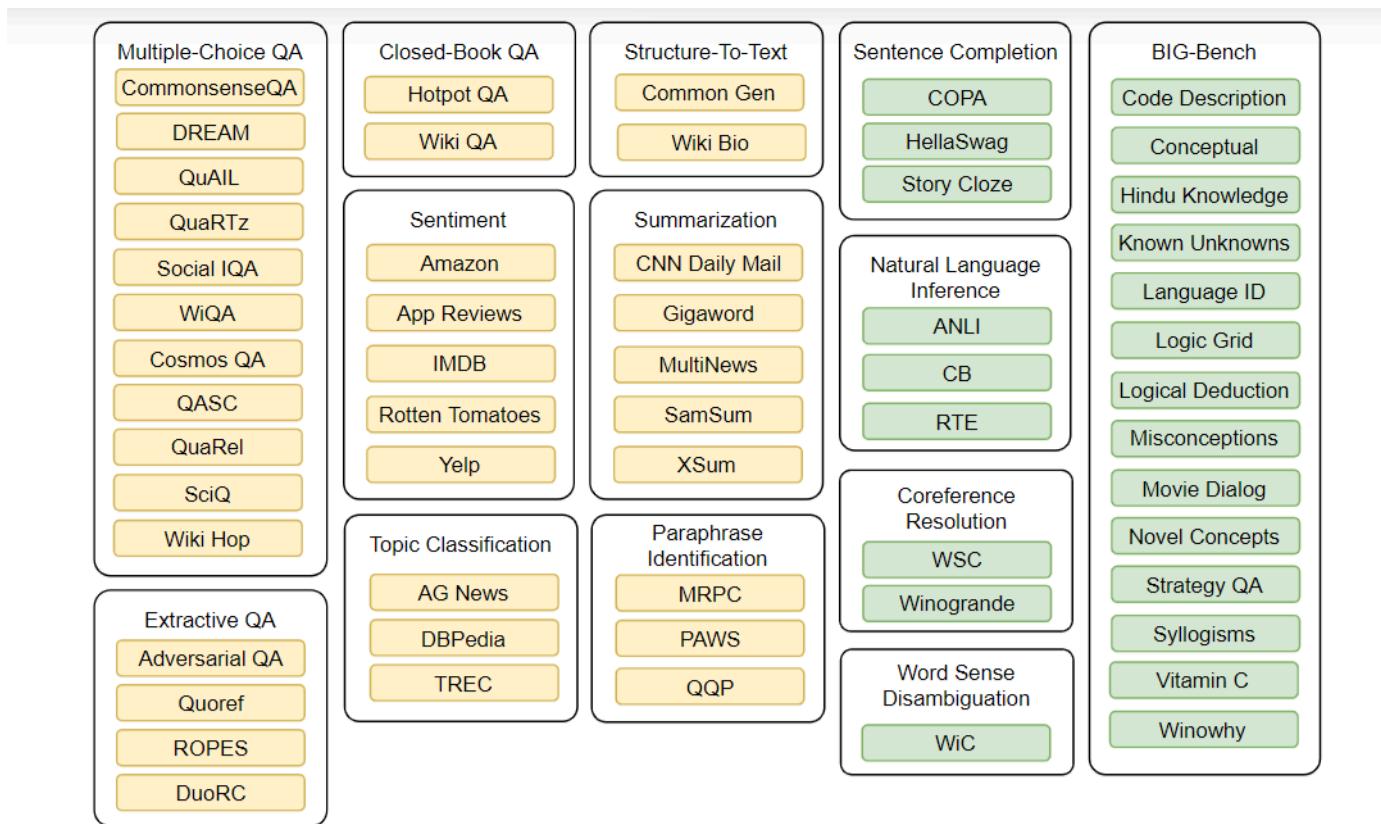


Figure 2: T0 datasets and task taxonomy. (T0+ and T0++ are trained on additional datasets. See Table 5 for the full list.) Color represents the level of supervision. Yellow datasets are in the training mixture. Green datasets are held out and represent tasks that were not seen during training. Hotpot QA is recast as closed-book QA due to long input length.

[2110.08207] Multitask Prompted Training Enables Zero-Shot Task Generalization

Caveat 4: multiple hypothesis testing

undisclosed private testing practices benefit a handful of providers who are able to test multiple variants before public release and retract scores if desired... At an extreme, we identify 27 private LLM variants tested by Meta in the lead-up to the Llama-4 release.

Singh (2025) The Leaderboard Illusion

Caveat 5: the 'general' benchmark paradigm

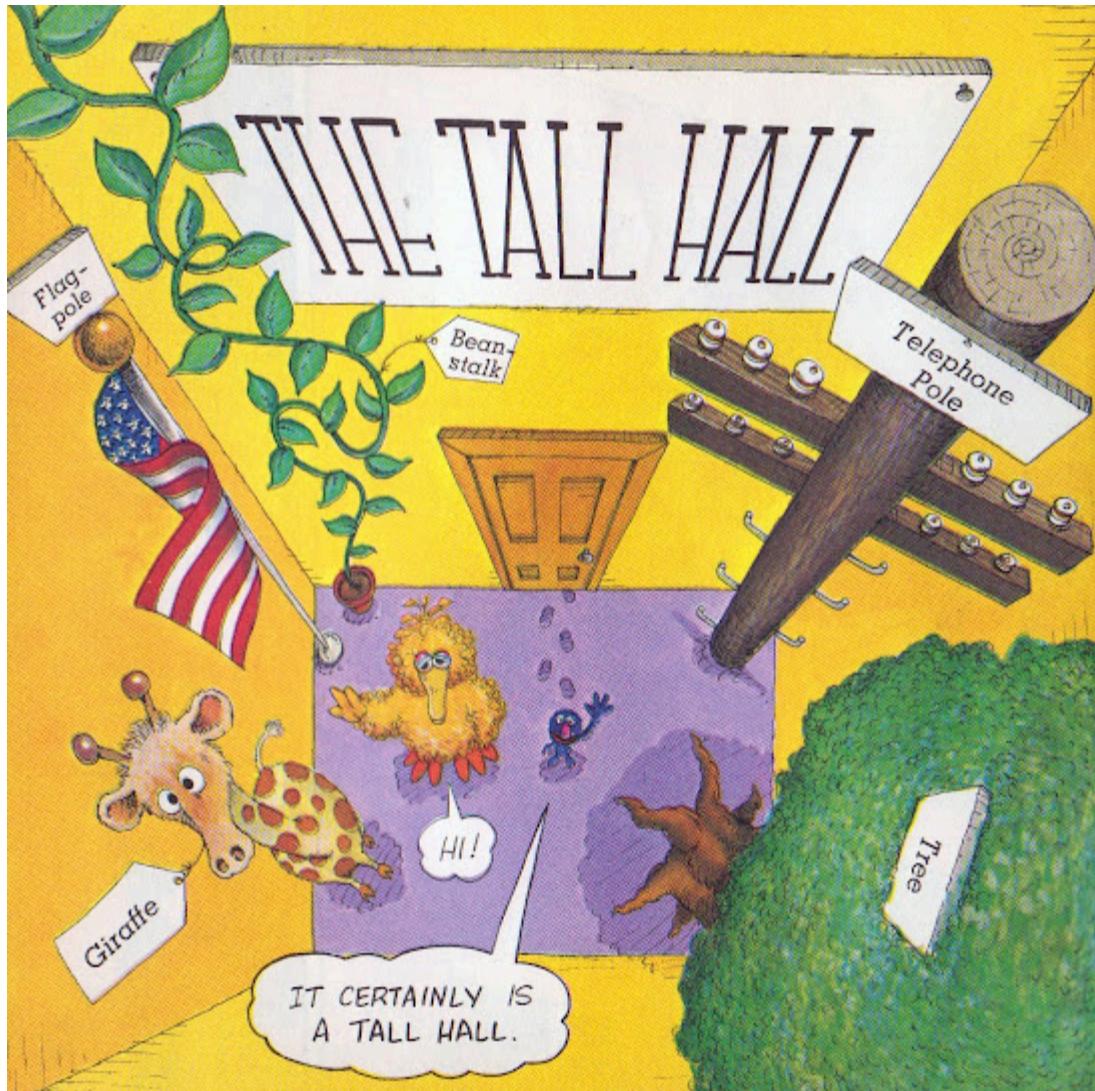
GLUE was designed to provide a general-purpose evaluation of language understanding that covers a range of training data volumes, task genres, and task formulations.

Wang et al. (2019) SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems

Are the 'general' benchmarks general?

If the so-called “general” benchmarks were legitimate tests of progress towards general artificial cognitive abilities, we would expect the tasks they embody to be chosen systematically, or with reference to specific theories of the cognitive abilities they model. Instead, what we observe looks more like samples of convenience...

Can a specific benchmark guarantee 'generality'?



Raji et al. (2021) [AI and the Everything in the Whole Wide World Benchmark](#), Image: [Grover and the Everything in the Whole Wide World Museum](#)

What would a principled sample even look like?

Case study: QuAIL dataset (multi-choice QA)

The Bear (Michael E. Shea)

The air exploded in a flash of bone and steel and blood. The clash of metal rang through the forest. An arrow pierced through the darkness, its barbed head tearing through flesh and muscle. A roar echoed off of the mountains far to the west. A cry broke through soon after. Then silence.

Char stood over a pile of black fur and red blood. He held a curved sword, jagged half way down the wide blade and hilted in bone. He held a large thick bow in the other. Lorfel and Ranur stood behind him, panting. Lorfel, a short man of twenty six held a large axe in both hands and still prepared to swing it hard. Ranur, the largest of the three held a pike in one hand, its tip hanging low towards the ground. He buried his other hand in his gray tunic.

"Did it get either of you?" Char's voice rasped low in the silence of the night.

"No" Lorfel said. He planted his axe head on the ground with a thud and leaned on the tall handle. There was a pause. Char turned towards Ranur.

"Are you hurt?"

"Mm...My hand." Ranur took his hand out of his tunic. Moonlight gleamed red off of the ragged wound. Char thought he saw a glimmer of bone.

"Did he claw you or bite you?" Char's voice held an urgency that set both Lorfel and Ranur on edge.

Ranur paused and then spoke low. "He bit me."

Char picked Lorfel and Ranur as his hunting partners for their speed and sharpness in battle. They had hunted beasts of the deep woods all of their lives. They hunted the beasts that hunted men. They all knew the risks of battling such creatures. The old man dropped his curved sword, drew his bow, and fired. The arrow hammered into Ranur's chest, burying itself in his heart. Lorfel saw the gleaming arrow head sticking almost a foot out of his companion's back. Ranur fell face first to the ground.

Rogers et al. (2020) Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks

Is this set 'representative' of QA in general?

Text-based questions

Q: When did the roar happen?

- A. before the cry
- B. after the silence
- C. NEI*
- D. when Char was speaking

Temporal order

* Not enough information to answer this question

Q: Who bit Ranur?

- A. the beast
- B. Lorfel
- C. Char
- D. NEI

Coreference

Q: What color was the beast's fur?

- A. brown
- B. NEI
- C. black
- D. red

Factual questions

Unanswerable questions

Q: What was done with Ranur's body?

- A. burned to avoid spreading disease
- B. left abandoned along with the beasts' corpse
- C. buried in the ground
- D. NEI

World knowledge questions

Q: Why was there blood?

- A. because Char shot something
- B. NEI
- C. because Lorfel had an axe
- D. because Char had a sword

Causality

Q: After the end of this text, Ranur is:

- A. standing up
- B. NEI
- C. on the ground
- D. in the sky

Subsequent state

Q: Ranur probably died:

- A. a month later
- B. instantly
- C. NEI
- D. a year later

Event duration

Q: What is probably true about the beast's bite?

- A. it is harmless
- B. it is extremely dangerous
- C. NEI
- D. it helps people

Properties

Q: Who was concerned about his companions' injuries?

- A. NEI
- B. Char
- C. Lorfel
- D. Ranur

Belief states



Talk to your neighbor!

- in your recent project, what phenomenon are you testing?
- how do you know that the dataset encapsulates that?
- what is your selection of models, and why are these models the ones to test?
- how do you know that the model was/wasn't contaminated?

HOW WELL DO LLMS PROCESS LONG CONTEXTS?

Ways to increase the processed context length

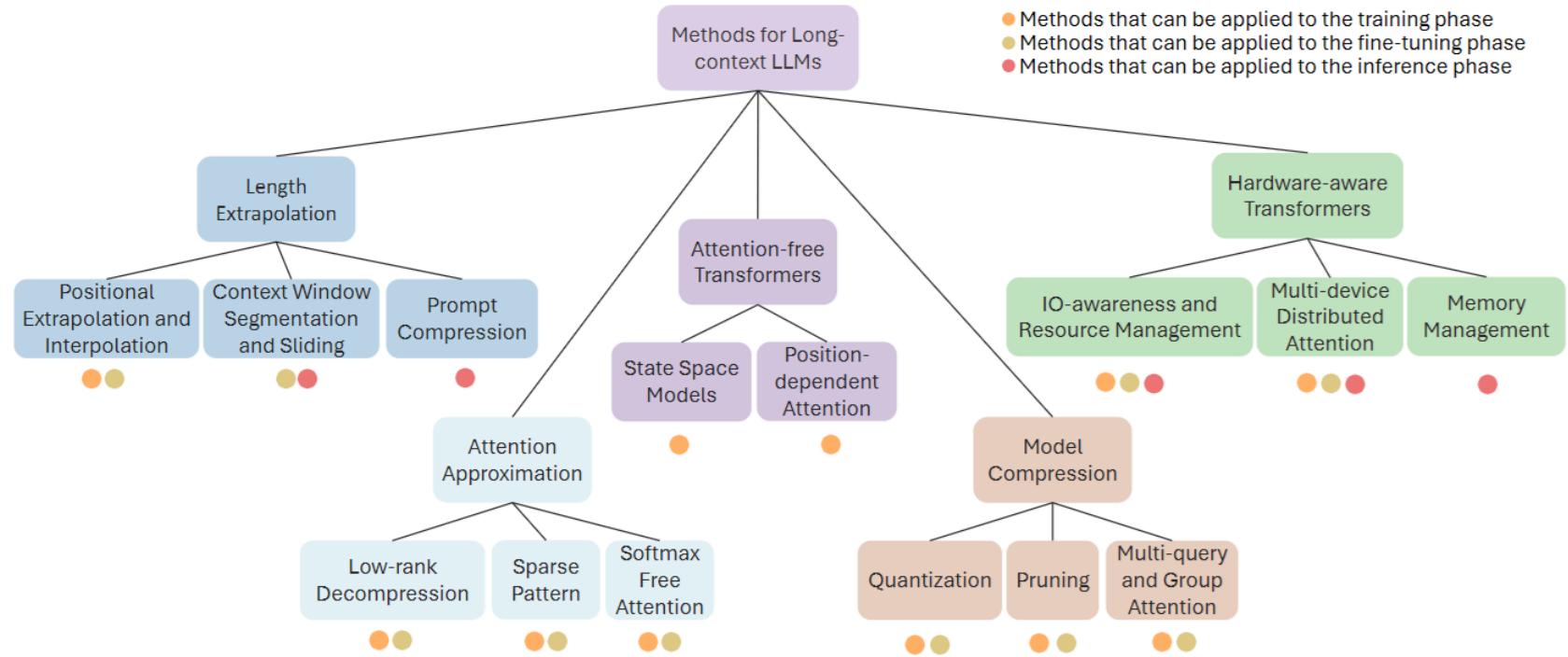
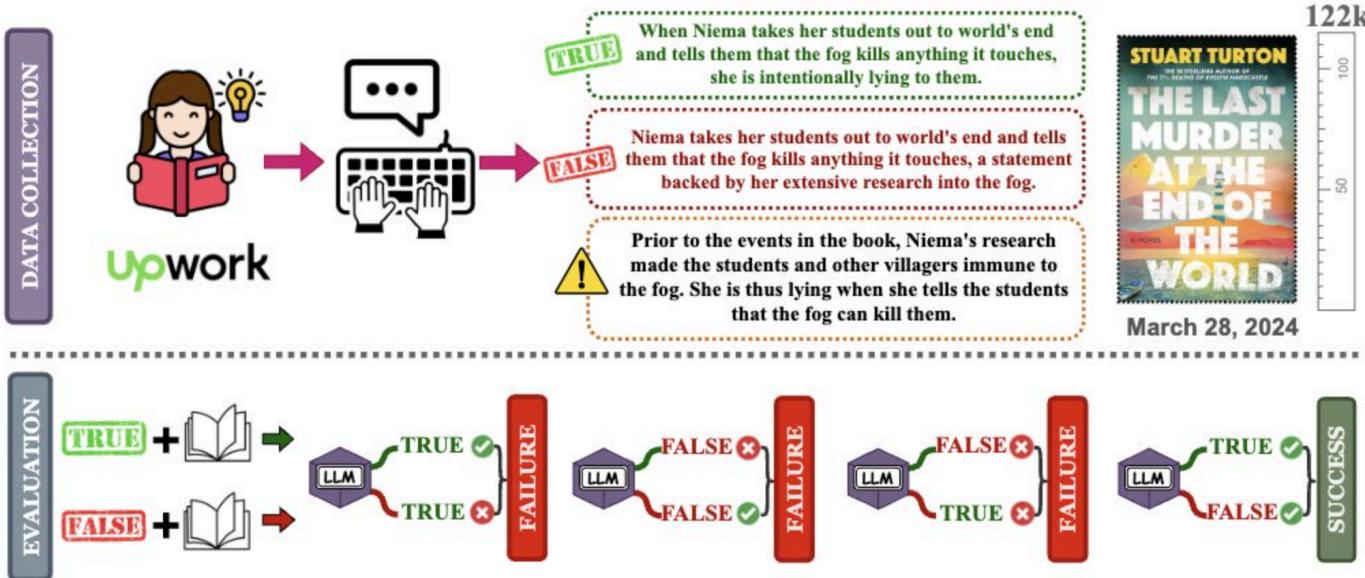


Figure 1: Taxonomy of Long-context LLM literature, which includes five distinct sections: length extrapolation, attention approximation, attention-free transformers, model compression, and hardware-aware transformers. We also establish connections between the methodologies and their related applicability scenarios. Some entail training a new model from scratch, others involve fine-tuning pre-trained models, and some implement over inference without any updates of hyper-parameters.

Wang et al. (2024) [Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models](#)

⚠ Accuracy problems when processing long contexts



MODEL	PAIR ACC _(correct/total)
GPT-4O	55.8 (344/617)
GPT-4-TURBO	40.2 (248/617)
CLAUDE-3-OPUS	49.4 (463/937)
CLAUDE-3.5-SONNET	41.0 (384/937)
GEMINI PRO 1.5	48.1 (247/514)
GEMINI FLASH 1.5	34.2 (176/515)

BM25+GPT-4O ($k=5$)	28.2 (282/1001)
BM25+GPT-4O ($k=25$)	44.1 (441/1001)
BM25+GPT-4O ($k=50$)	49.7 (497/1001)
RANDOM	25.0 (250/1001)

[2406.16264] One Thousand and One Pairs: A "novel" challenge for long-context language models

⚠ Accuracy problems when processing long contexts

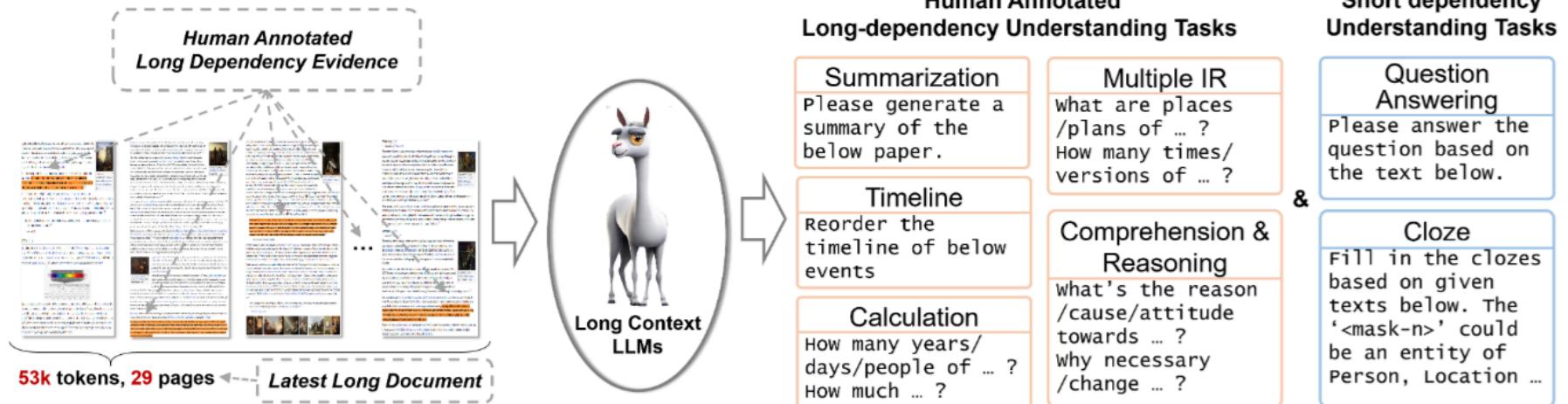


Figure 1: The LooGLE benchmark for long context understanding.

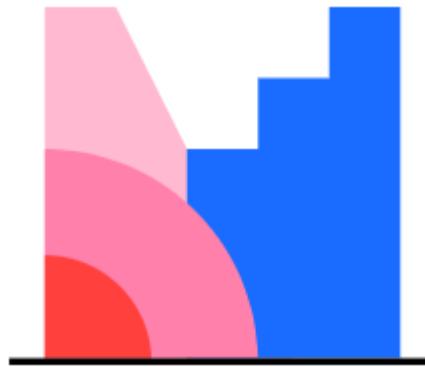
Models	Information & retrieval	Timeline & reorder	Calculation	Comprehension & reasoning
GPT4-32k	33.26	26.43	22.30	44.20
GPT4-8k	26.59	20.61	16.31	34.42
GPT3.5-turbo-16k	24.05	20.88	13.49	32.10
LlamaIndex	19.38	17.23	11.43	29.53
ChatGLM2-6B-32k	11.38	10.77	8.45	10.95
LongLLaMa-3B-Instruct	15.73	8.87	8.87	21.29
RWKV-4-14B-raven	5.73	4.76	2.08	6.52
LLaMA2-7B-32K-Instruct	2.23	1.36	1.39	2.67

Table 7: Individual task results of long dependency QAs

[LooGLE: Can Long-Context Language Models Understand Long Contexts?](#) (Li et al., ACL 2024)

DO LLMS HAVE EMERGENT PROPERTIES?

What do YOU think?



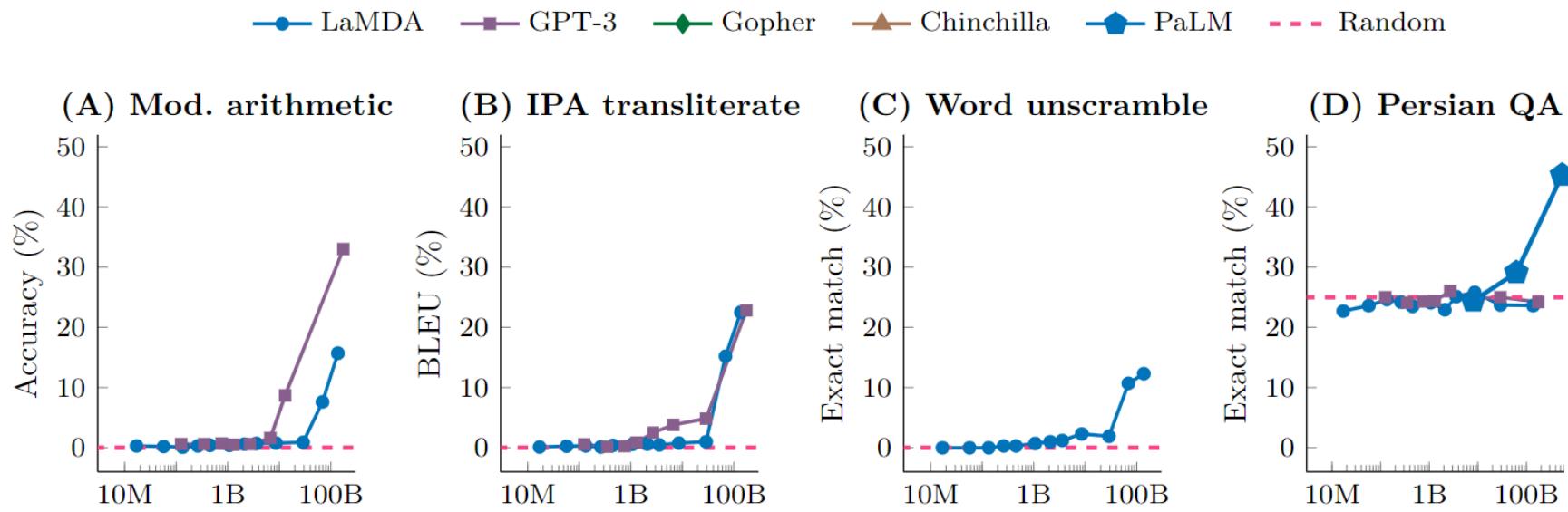


What do we even mean by 'emergent properties'?

Rogers, Luccioni (ICML 2024) [Position: Key Claims in LLM Research Have a Long Tail of Footnotes](#)

'Emergence': definition 1

🤔 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."



Wei et al. (2022) [Emergent Abilities of Large Language Models](#)

Discussion: definition 1.

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

If data similar to test data was included in training,
'emergence' is to be expected in larger models!

'Emergence': definition 2

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

Schaeffer et al. (NeurIPS 2023 oral) show that the observed sharpness is an artifact of non-continuous evaluation metrics

Schaeffer et al. (2023) [Are Emergent Abilities of Large Language Models a Mirage?](#)

'Emergence': definition 3

 *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 "machine learning"?

Deshpande et al. (2023) [Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.](#)

A twist on definition 3: 'emergence during training'

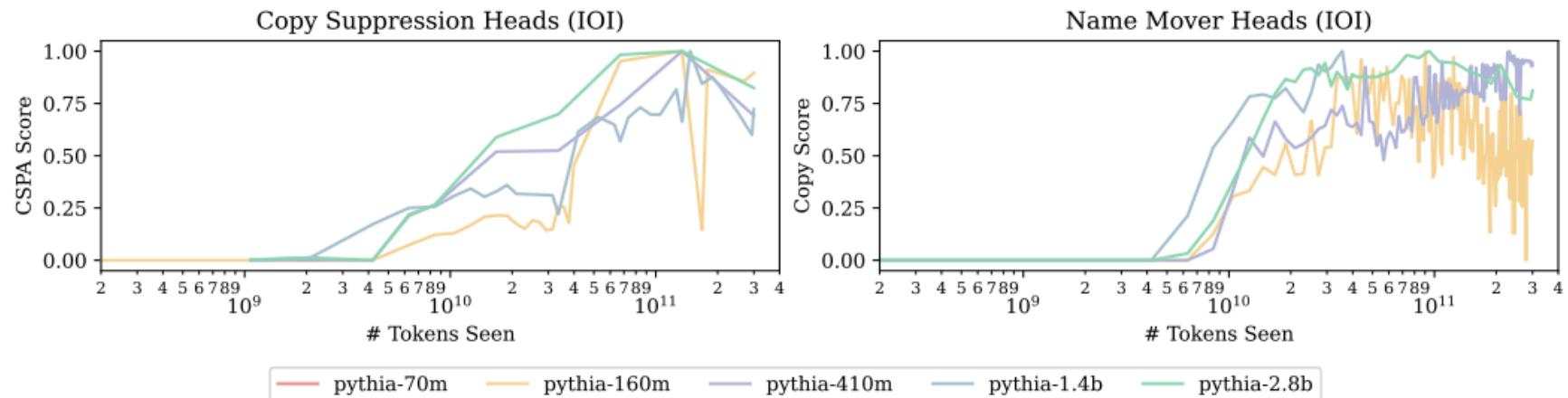


Figure 2: The development of components relevant to IOI and Greater-Than, across models and time. Each line indicates the degree to which attention heads in the circuit at each timestep exhibit the relevant component behavior. The timesteps at which component behavior emerges parallel those at which task performance emerges in Figure 1.

can we just say "training dynamics"?

'Emergence': definition 4

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

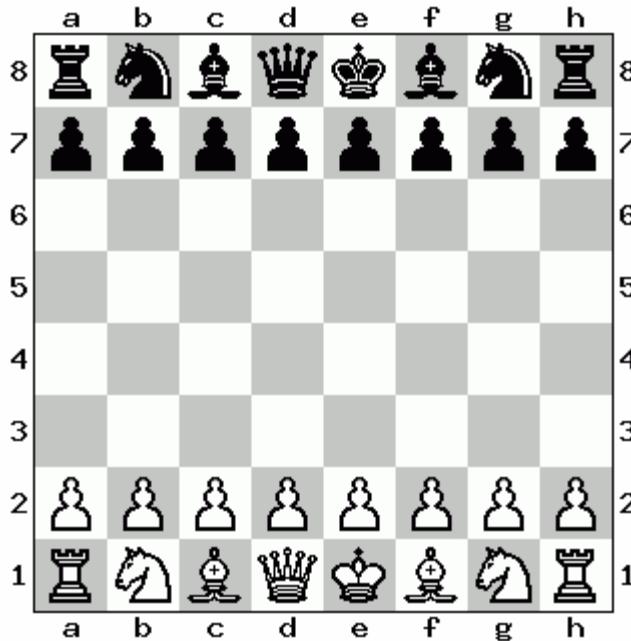
- cannot show this without analysis of pre-training data!
- even for "open" models, no methodology so far to do analysis of supporting evidence beyond the obvious memorization

A twist on definition 4

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?

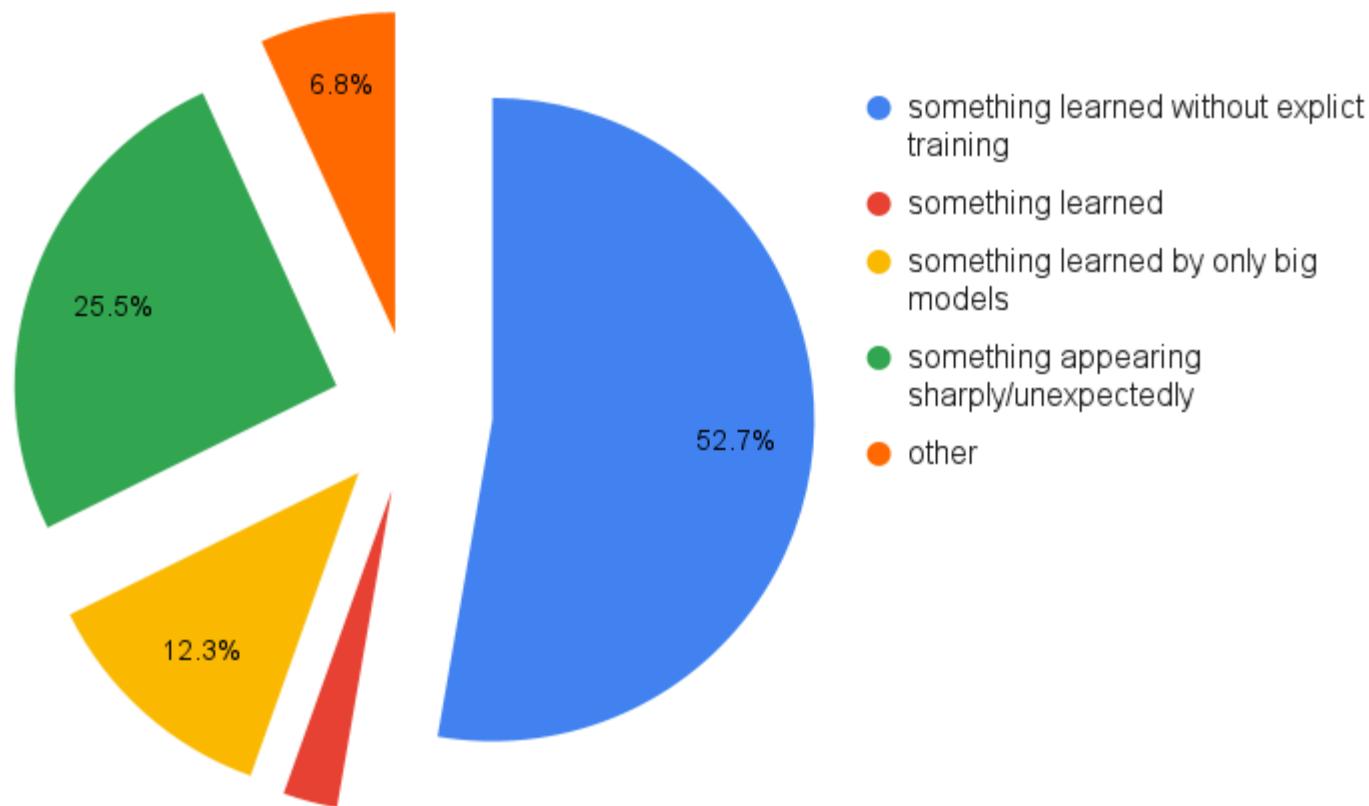


Training LLMs is an expensive way to discover... that the Internet contains chess data?

chatgpt: black, stockfish: white. source: [r/AnarchyChess](#)

For most NLP researchers, 'emergence' ~ 'generalization'!

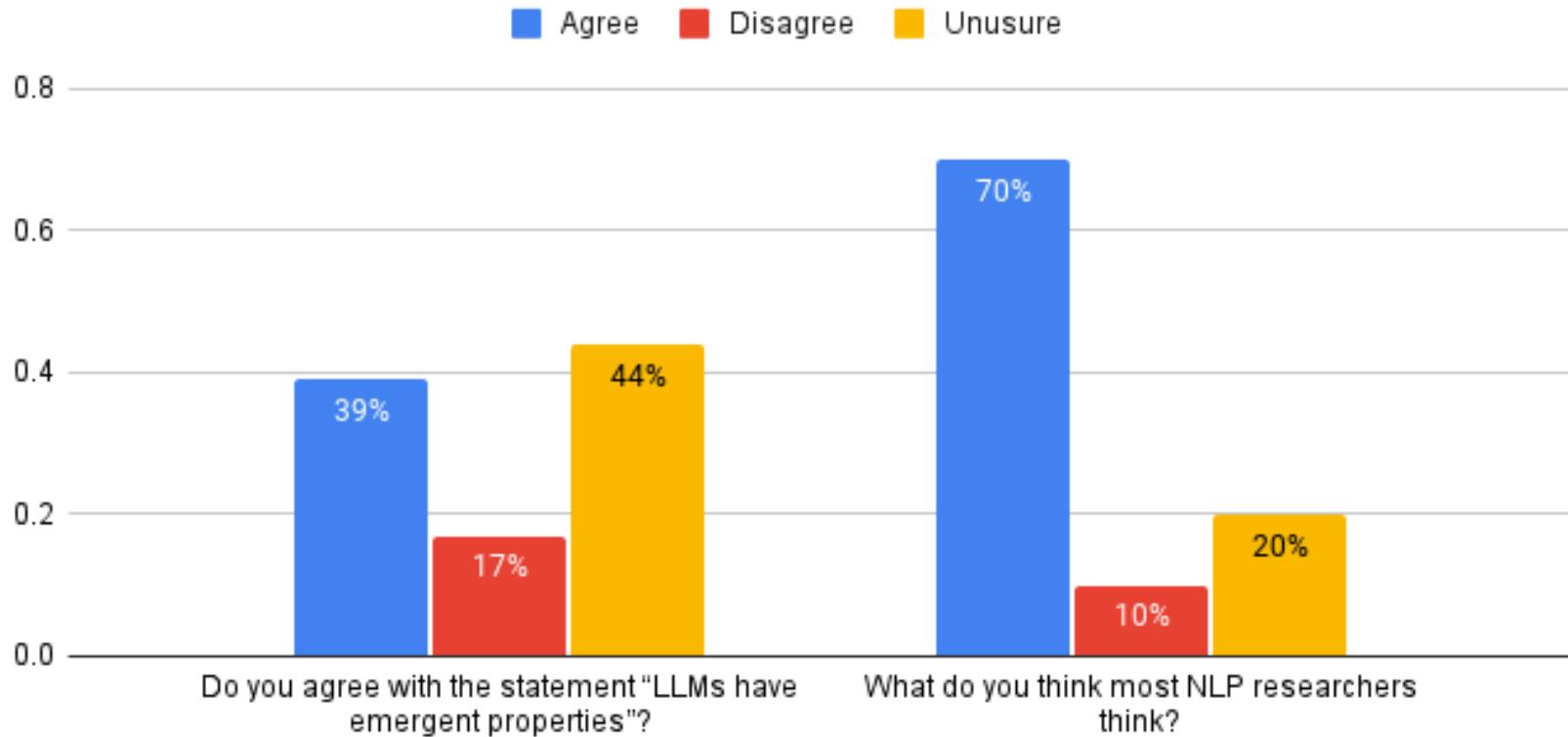
How do you define 'emergent properties' with respect to LLMs?



Survey results of 220 NLP researchers & PhD students: <https://hackingsemantics.xyz/2024/emergence/>

Most NLP researchers don't believe in 'emergence'!

Do you agree with the statement "LLMs have emergent properties"?



Survey results of 220 NLP researchers & PhD students: <https://hackingsemantics.xyz/2024/emergence/>

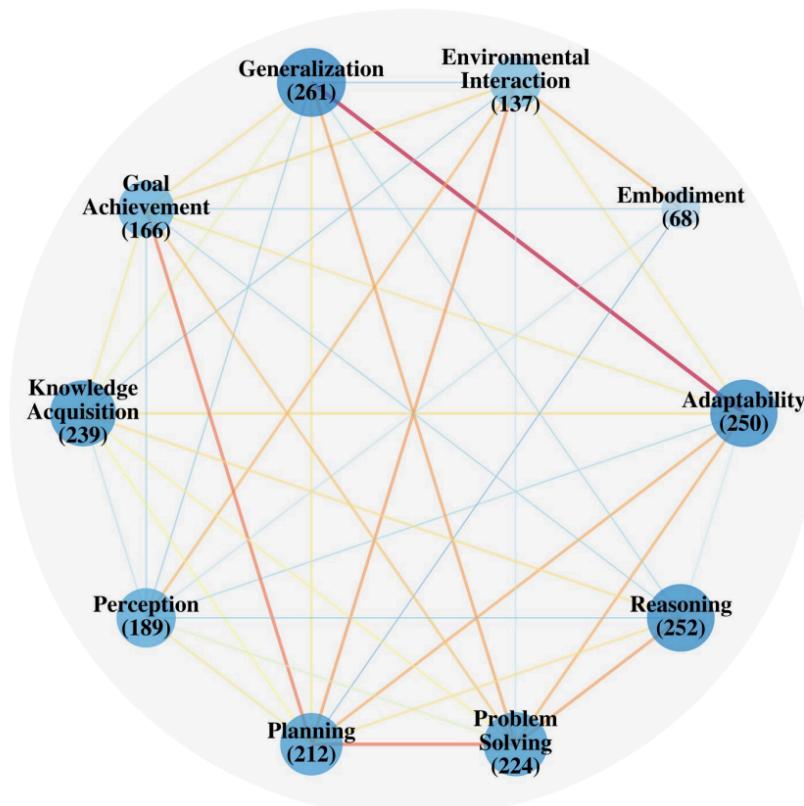


Figure 1: Correlation between criteria that the survey respondents selected as relevant for their notion of “intelligence”. Darker edges indicate stronger correlations, larger nodes indicate higher relevance. Only edges with $\phi > |0.1|$ are shown.

Fun fact:
Researchers also
associate
'generalization' with
'intelligence'

Emergent properties in philosophy

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

- Weight patterns can be viewed as "functional realization" of what they're supposed to model
- "emergence" is still equivalent to "machine learning"?

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

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

Lu et al. (2023) Are Emergent Abilities in Large Language Models just In-Context Learning?

Conclusions of Lu et al.

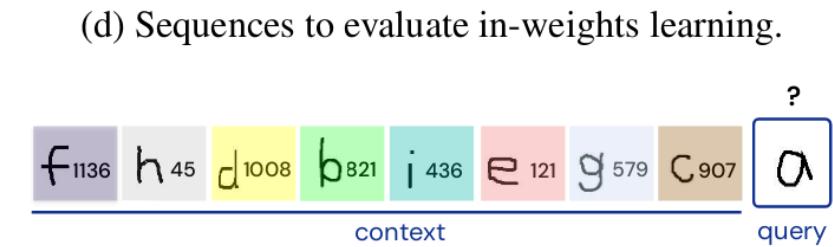
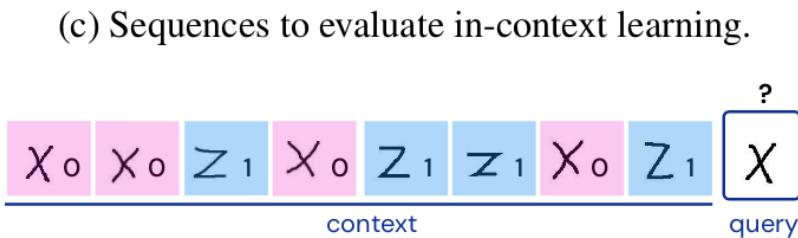
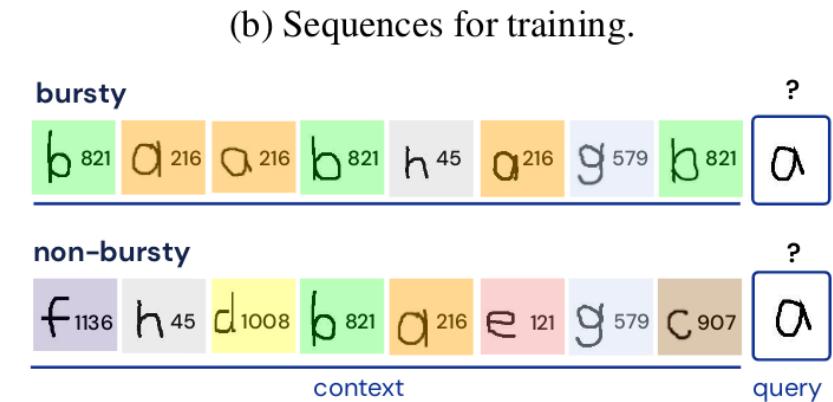
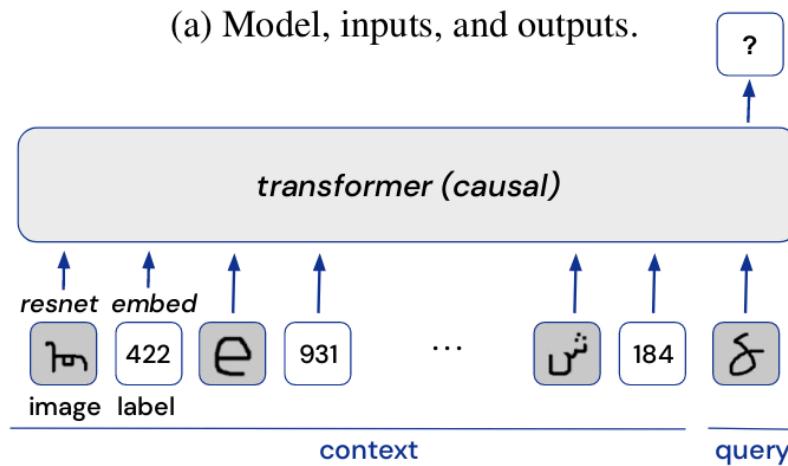
- **nearly all emergent LLM functionalities are attributable to in-context learning!**
- **instruction tuning allows for better use of in-context learning, rather than independently causes emergent functionalities**

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.

Lu et al. (2023) Are Emergent Abilities in Large Language Models just In-Context Learning?

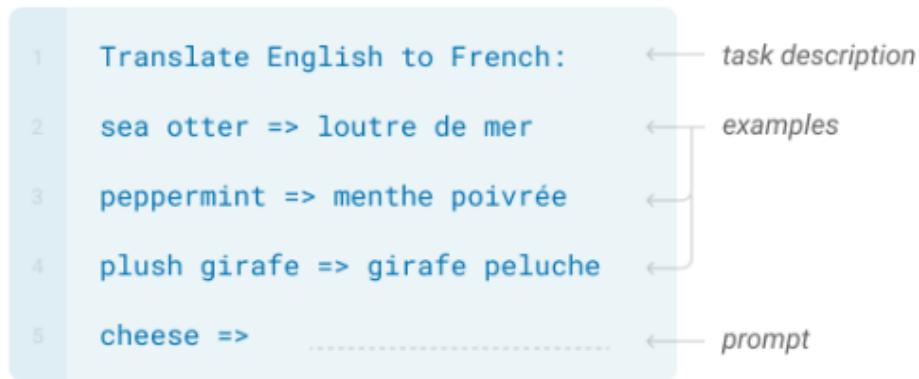
In-context-learning is driven by training data



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

Few-shot

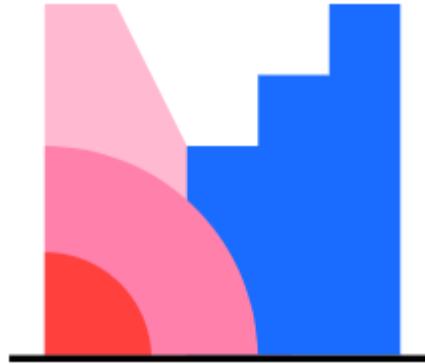
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



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

Brown et al. (2020) [Language Models are Few-Shot Learners](#)

What's your take?



Takeaways from LLM factuality discussion

- ~~Anna hates LLMs~~
- ~~LLMs are useless~~
- ~~LLMs don't model meaning at all~~
- As any tool, LLMs can be useful *when their utility is appropriately scoped*

? Are they appropriately scoped now?

Why we need to be careful: media

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/164881111111111111)

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



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



Why we need to be careful: politicians



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.

Why we need to be careful: our own wishful thinking

Although it is not appropriate to apply LLMs directly for extracting arguments, we believe that the emergence capabilities of LLMs hold promise for D-EAE models to model complex implicit associations in events

Takeaways

As researchers, we need to be more careful with claims & terminology!

- 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](#)



Thank you!

📢 🇩🇰 PhD position on expertise recommendations @ ITU

📢 🇩🇰 postdoc on real-world LLM eval @AAU (with Roman Jurowetzki & me)!

✉️ arog@itu.dk

[@annarogers.bsky.social](https://annarogers.bsky.social)

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

