Large Language Models and Factuality

Part 3: Facts on LLMs

AthensNLP

Athens, September 24 2024

Anna Rogers

📢 🇩🇰 **open PhD and postdoc positions!**

IT UNIVERSITY OF COPENHAGEN

To what extent do you agree with the following statements?

Part 3. Facts *on* LLMs

- LLMs are SOTA
- LLMs are robust
- LLMs are General Purpose Technologies
- LLMs are good at processing long contexts
- LLMs have emergent properties

LLMS ARE SOTA

Are you likely to get a higher performance on X with a bigger Transformer?

Yes, but...

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

Caveat: 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

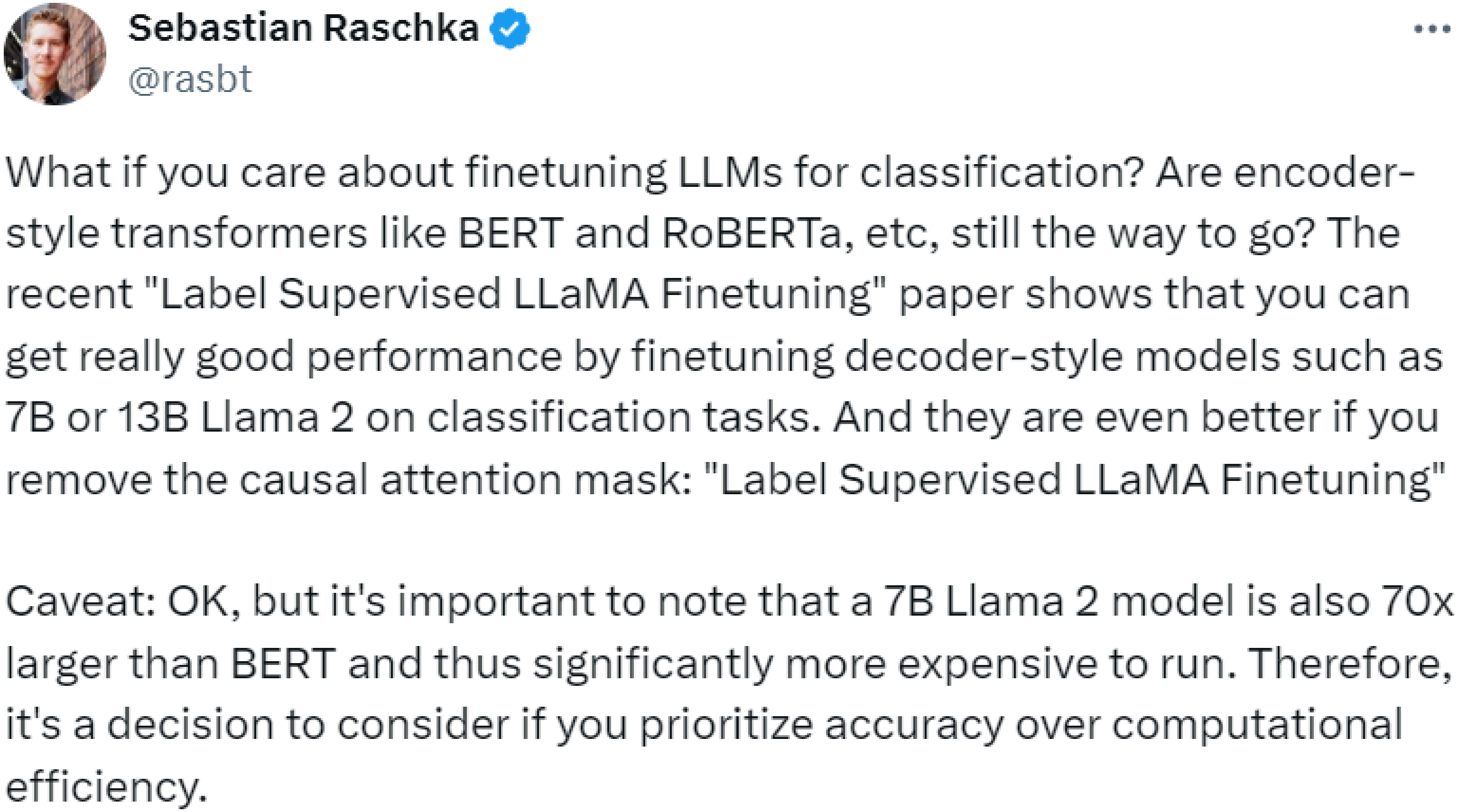
https://super.gluebenchmark.com/leaderboard/

Caveat: *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: classification tasks



Sebastian Raschka
@rasbt

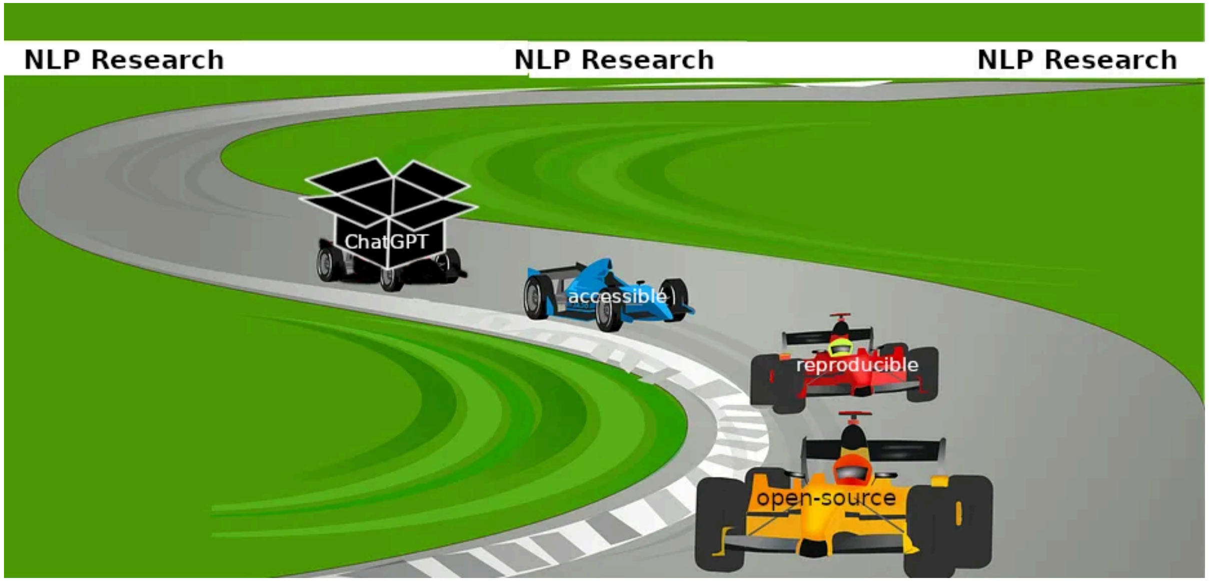
What if you care about finetuning LLMs for classification? Are encoder-style transformers like BERT and RoBERTa, etc, still the way to go? The recent "Label Supervised LLaMA Finetuning" paper shows that you can get really good performance by finetuning decoder-style models such as 7B or 13B Llama 2 on classification tasks. And they are even better if you remove the causal attention mask: "Label Supervised LLaMA Finetuning"

Caveat: OK, but it's important to note that a 7B Llama 2 model is also 70x larger than BERT and thus significantly more expensive to run. Therefore, it's a decision to consider if you prioritize accuracy over computational efficiency.

Sebastoan Raschka's post, Label Supervised LlaMA Finetuning

🤔 'Closed' LLMs are SOTA

That which is not open and reasonably reproducible cannot be considered a requisite baseline.



Rogers A. (2023) Closed AI Models Make Bad Baselines

LLMS ARE ROBUST

PROMPT SENSITIVITY

Caveat: prompt sensitivity

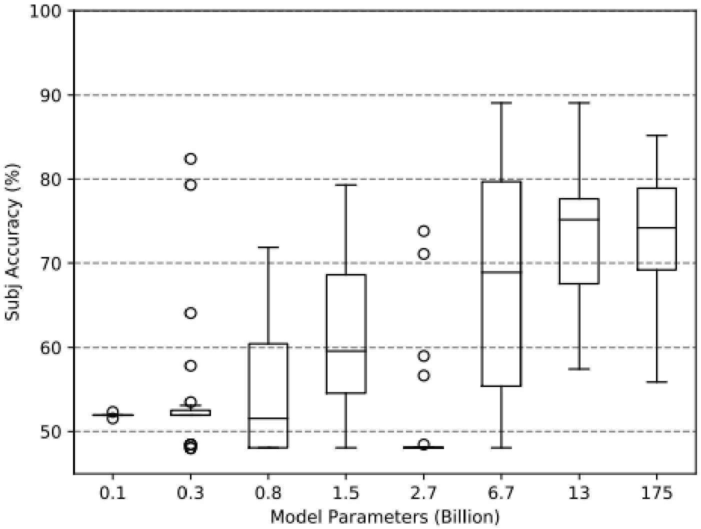


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.

- the order of samples and prompt template make a lot of difference!
- in model-API settings, your carefully-tuned prompt might stop working after a model update

Lu et al. (2022) Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity

Further caveat: label bias

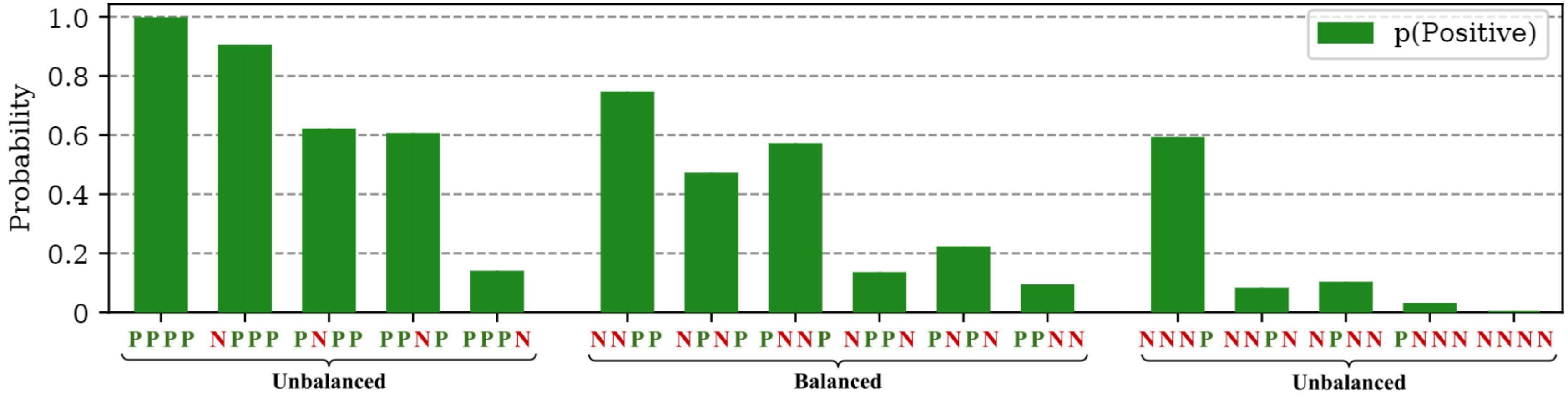
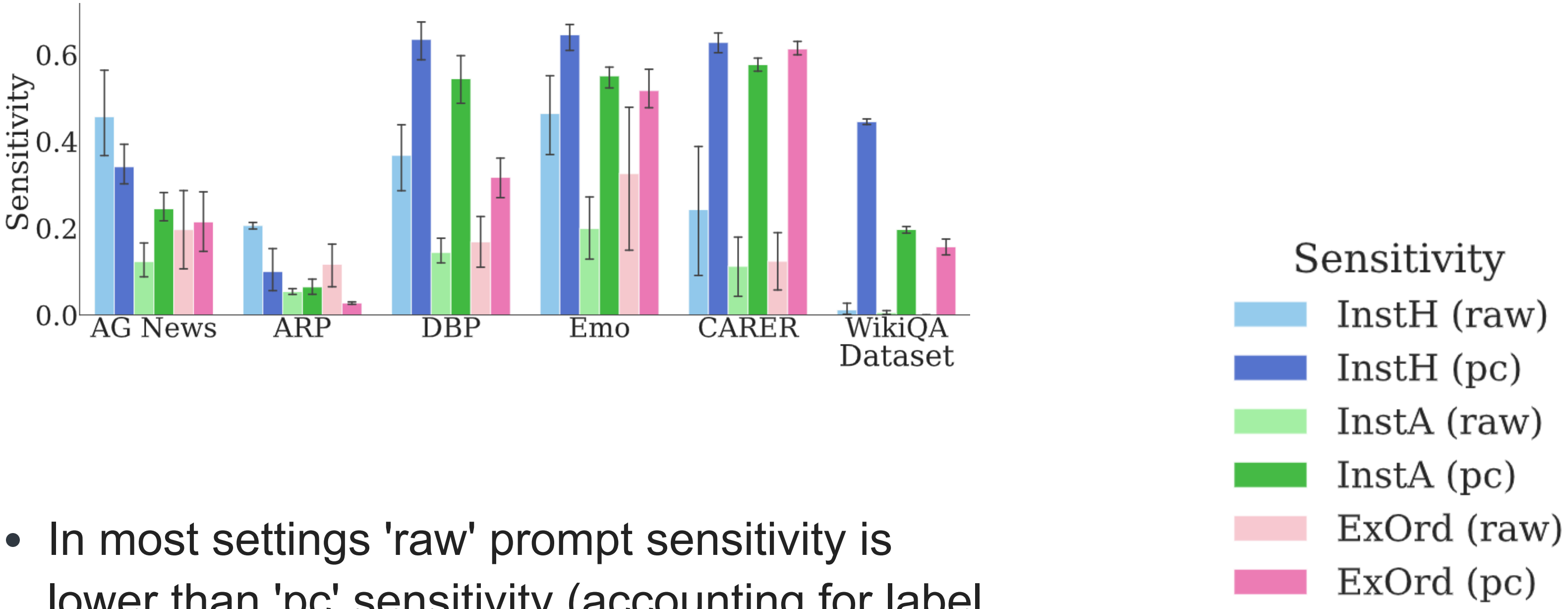


Figure 4. **Majority label and recency biases** cause GPT-3 to become biased towards certain answers and help to explain the high variance across different examples and orderings. Above, we use 4-shot SST-2 with prompts that have different class balances and permutations, e.g., [P P N N] indicates two positive training examples and then two negative. We plot how often GPT-3 2.7B predicts Positive on the balanced validation set. When the prompt is unbalanced, the predictions are unbalanced (*majority label bias*). In addition, balanced prompts that have one class repeated near the end, e.g., end with two Negative examples, will have a bias towards that class (*recency bias*).

Zhao et al. (2021) Calibrate Before Use: Improving Few-shot Performance of Language Models

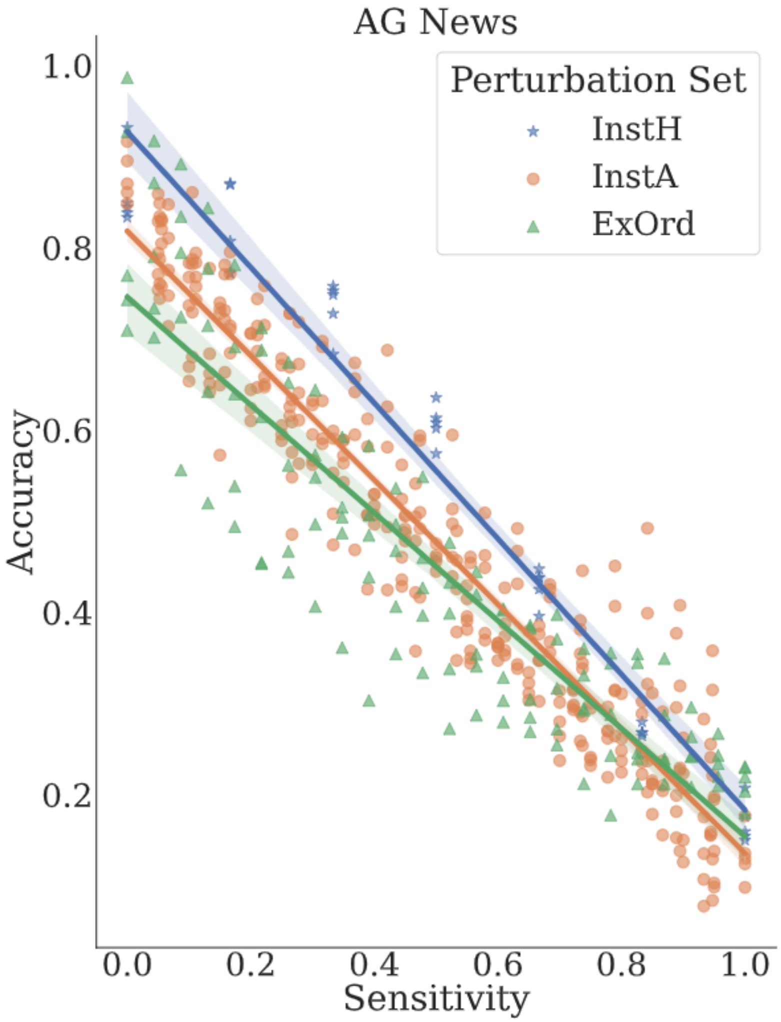
Prompt sensitivity is underestimated in evaluations!



- In most settings 'raw' prompt sensitivity is lower than 'pc' sensitivity (accounting for label bias)!
- adjusted sensitivity after mitigating label bias is up to 2.8x of the raw sensitivity!

Chen et al. (2023) On the Relation between Sensitivity and Accuracy in In-Context Learning

Prompt sensitivity is underestimated in evaluations!



Significant negative correlation between prediction sensitivity and accuracy (measured over examples with that sensitivity)

Chen et al. (2023) On the Relation between Sensitivity and Accuracy in In-Context Learning

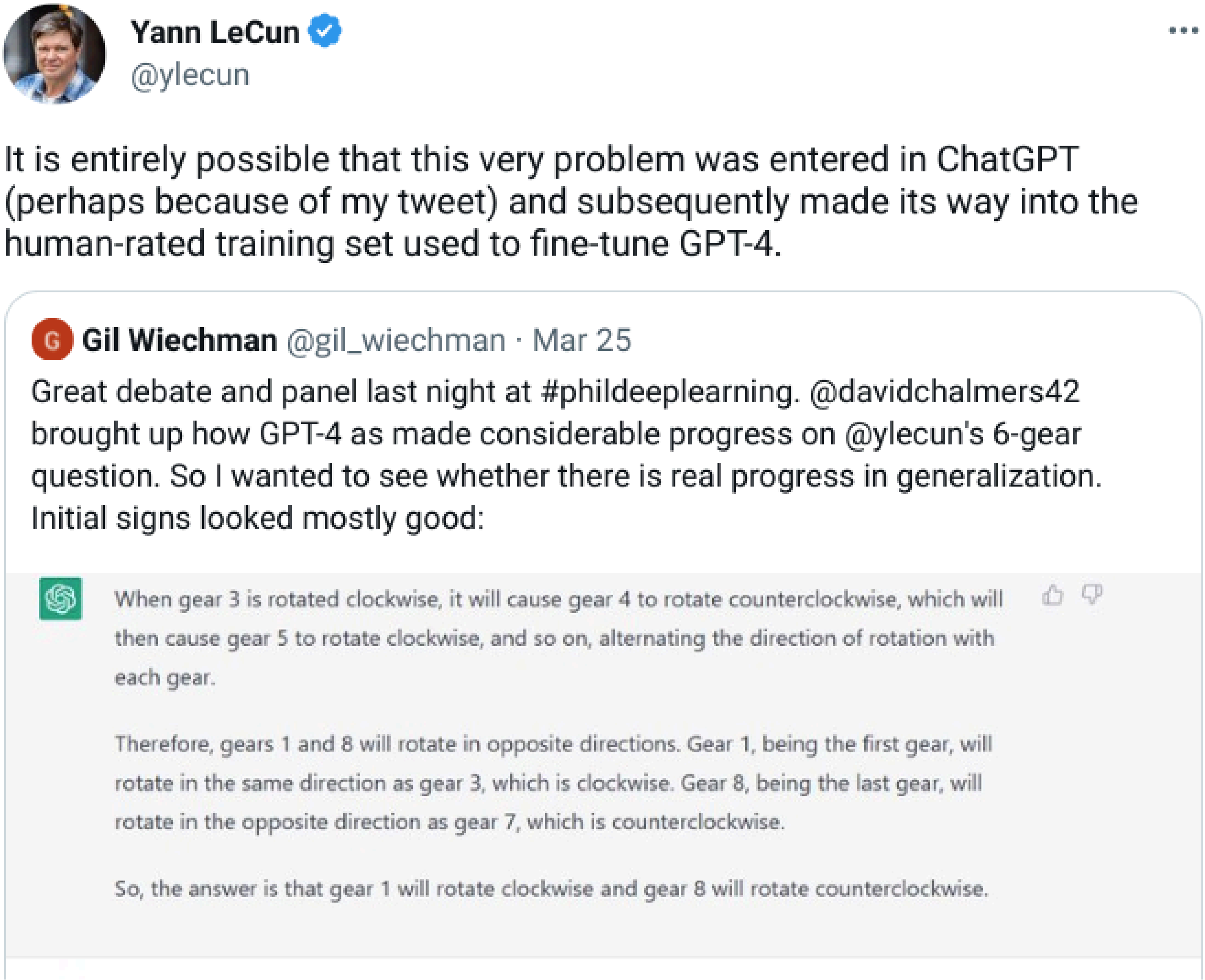
TEST DATA CONTAMINATION

Obvious contamination

LM contamination index

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

ChatGPT whack-a-mole



Yann LeCun @ylecun

It is entirely possible that this very problem was entered in ChatGPT (perhaps because of my tweet) and subsequently made its way into the human-rated training set used to fine-tune GPT-4.

Gil Wiechman @gil_wiechman · Mar 25

Great debate and panel last night at #phildeeplearning. @davidchalmers42 brought up how GPT-4 as made considerable progress on @ylecun's 6-gear question. So I wanted to see whether there is real progress in generalization. Initial signs looked mostly good:

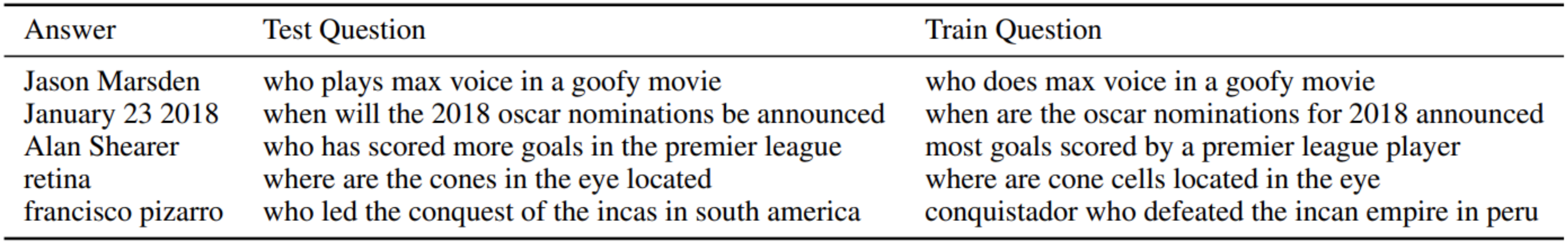
When gear 3 is rotated clockwise, it will cause gear 4 to rotate counterclockwise, which will then cause gear 5 to rotate clockwise, and so on, alternating the direction of rotation with each gear.

Therefore, gears 1 and 8 will rotate in opposite directions. Gear 1, being the first gear, will rotate in the same direction as gear 3, which is clockwise. Gear 8, being the last gear, will rotate in the opposite direction as gear 7, which is counterclockwise.

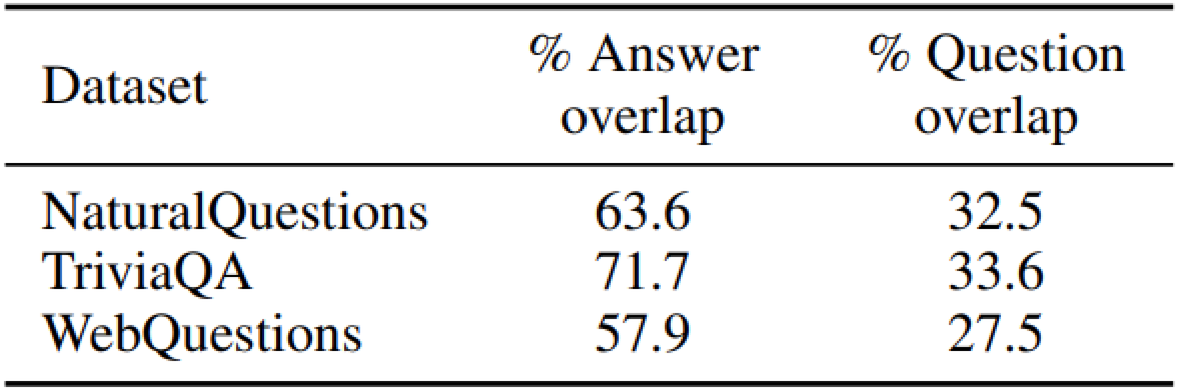
So, the answer is that gear 1 will rotate clockwise and gear 8 will rotate counterclockwise.

https://x.com/ylecun/status/1639685628722806786

Subtle train-test overlap



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

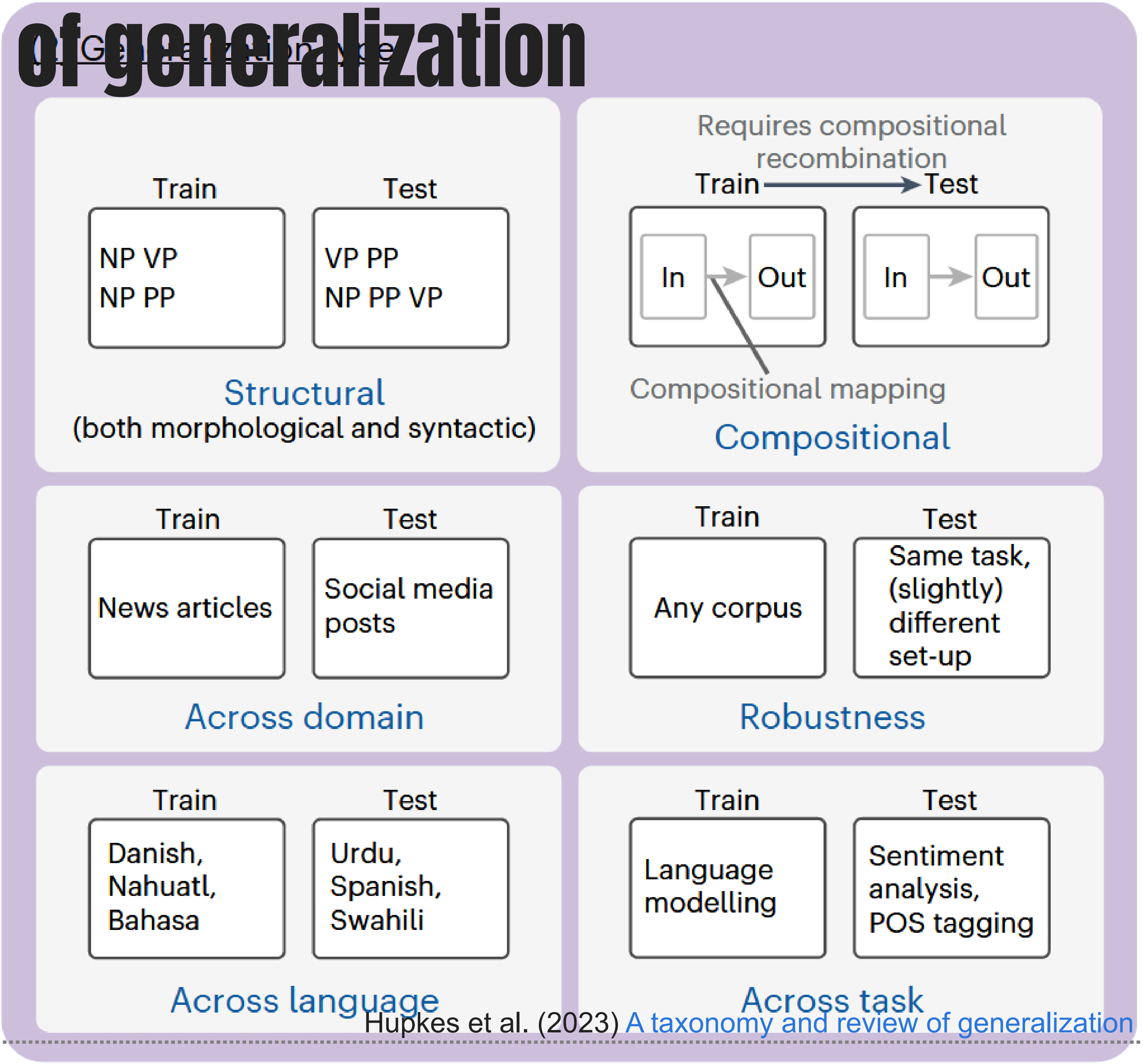
Lewis et al. (2021) Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets

Contamination vs generalization

~~Does model X generalize?~~

Does model X generalize to Y?

Types of generalization



Hupkes et al. (2023) A taxonomy and review of generalization research in NLP

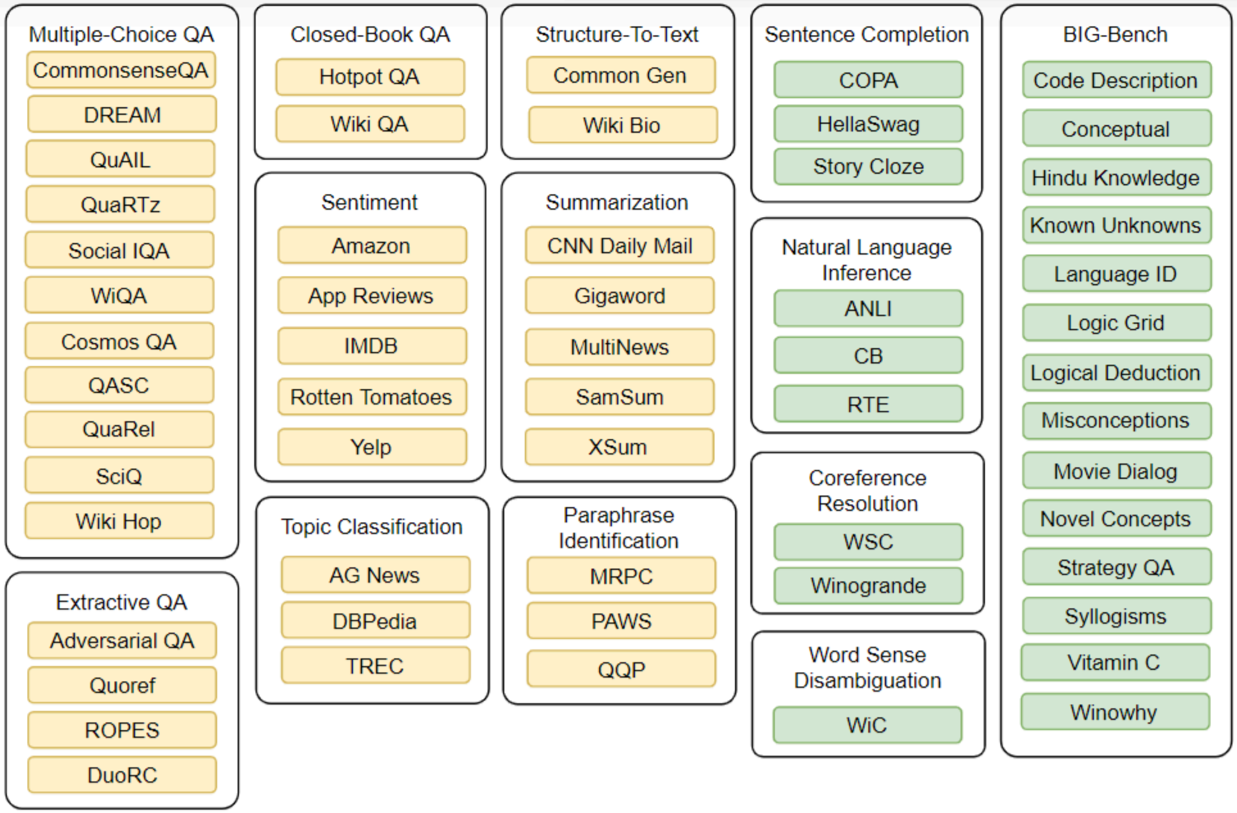
Example: task generalization claim

> *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*

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

Ouyang et al. (2022) Training language models to follow instructions with human feedback

Evaluation on generalizing to new tasks is inconclusive at best



Multiple-Choice QA
- CommonsenseQA
- DREAM
- QuAIL
- QuaRTz
- Social IQA
- WiQA
- Cosmos QA
- QASC
- QuaRel
- SciQ
- Wiki Hop

Extractive QA
- Adversarial QA
- Quoref
- ROPES
- DuoRC

Closed-Book QA
- Hotpot QA
- Wiki QA

Sentiment
- Amazon
- App Reviews
- IMDB
- Rotten Tomatoes
- Yelp

Topic Classification
- AG News
- DBPedia
- TREC

Structure-To-Text
- Common Gen
- Wiki Bio

Summarization
- CNN Daily Mail
- Gigaword
- MultiNews
- SamSum
- XSum

Paraphrase Identification
- MRPC
- PAWS
- QQP

Sentence Completion
- COPA
- HellaSwag
- Story Cloze

Natural Language Inference
- ANLI
- CB
- RTE

Coreference Resolution
- WSC
- Winogrande

Word Sense Disambiguation
- WiC

BIG-Bench
- Code Description
- Conceptual
- Hindu Knowledge
- Known Unknowns
- Language ID
- Logic Grid
- Logical Deduction
- Misconceptions
- Movie Dialog
- Novel Concepts
- Strategy QA
- Syllogisms
- Vitamin C
- Winowhy

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

Example: tikz generalization claim

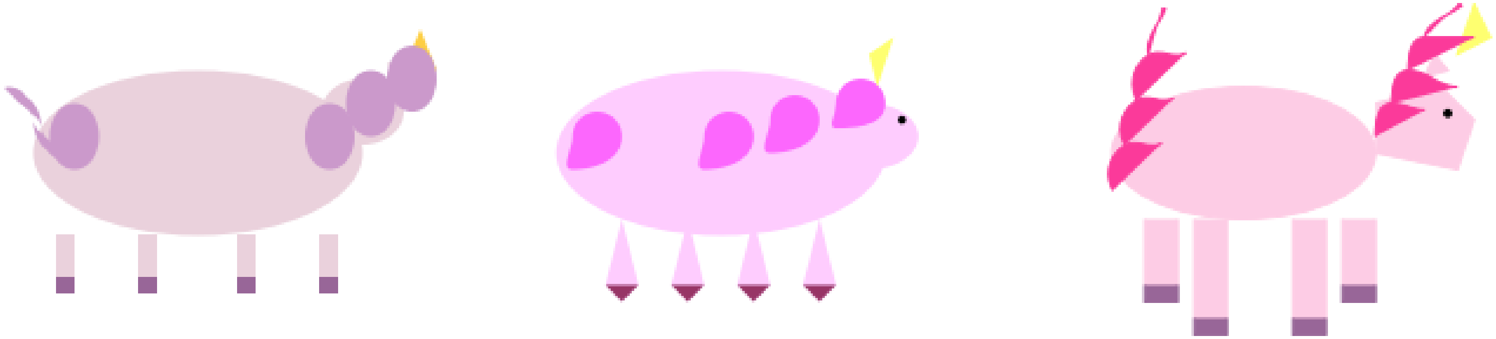
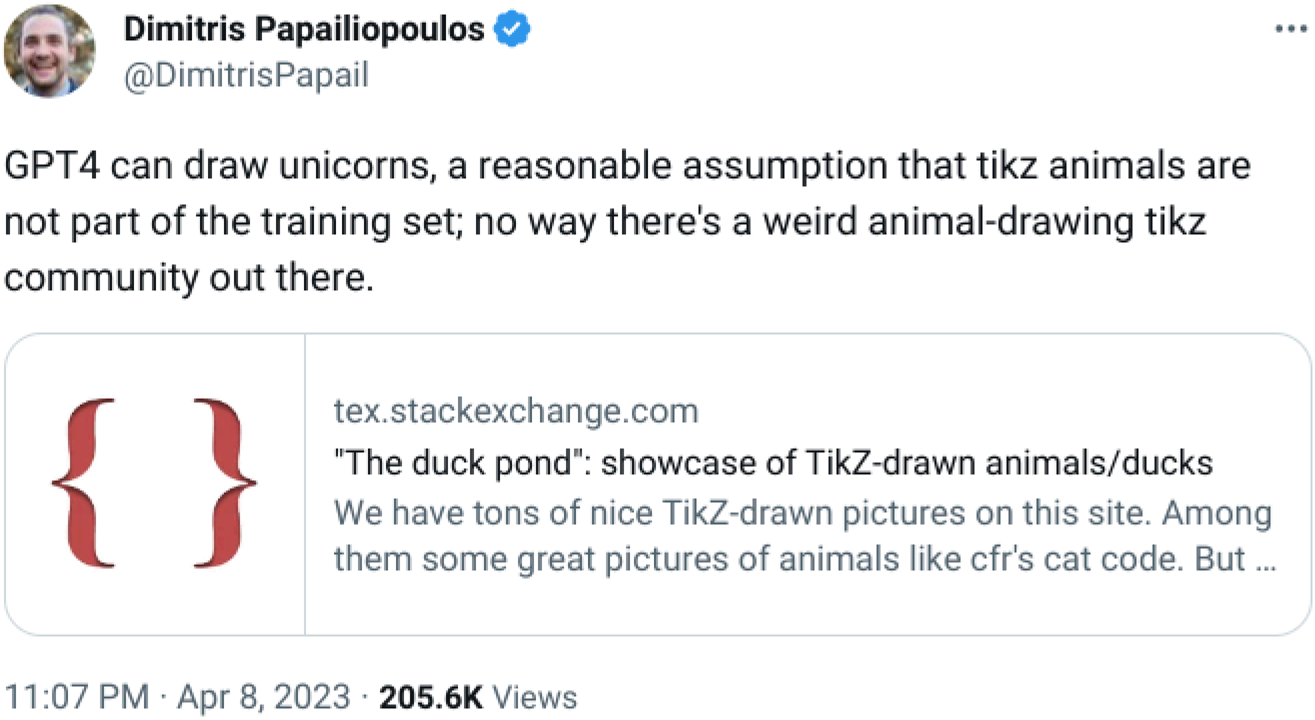


Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt "Draw a unicorn in TikZ". We can see a clear evolution in the sophistication of GPT-4's drawings.

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

However...



Dimitris Papailiopoulos
@DimitrisPapail

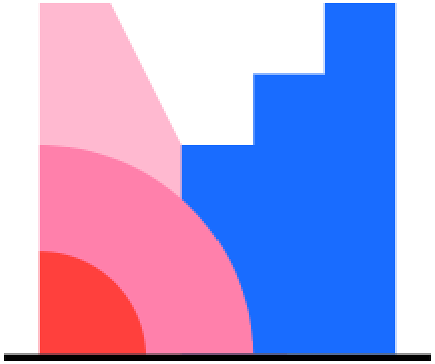
GPT4 can draw unicorns, a reasonable assumption that tikz animals are not part of the training set; no way there's a weird animal-drawing tikz community out there.

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

The scope of generalization of LLM-based chatbots?



Instruction tuning: example recent dataset

https://huggingface.co/datasets/allenai/tulu-v2-sft-mixture

(used to train OLMo-instruct)

Can a "general" NLP 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

Can a "general" NLP 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

LLMS ARE GPTS

GPTs are GPTs

... up to 49% of workers could have half or more of their tasks exposed to LLMs

Eloundou (2023) GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

General-purpose technology definition

1. it is a single, recognisable generic technology,
2. widely used across the economy,
3. it has many different uses
4. many spillover effects

A total of 24 technologies such as the wheel, the printing press and electricity are considered GPTs.

Lipsey et al. (2005) Economic transformations: general purpose technologies and long-term economic growth

Are GPTs GPTs?

it is a single, recognisable generic technology	❎
widely used across the economy	❎
has many different uses	✅
many spillover effects	❓

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

LLMS FOR LONG-CONTEXT PROCESSING

Ways to increase the processed context length

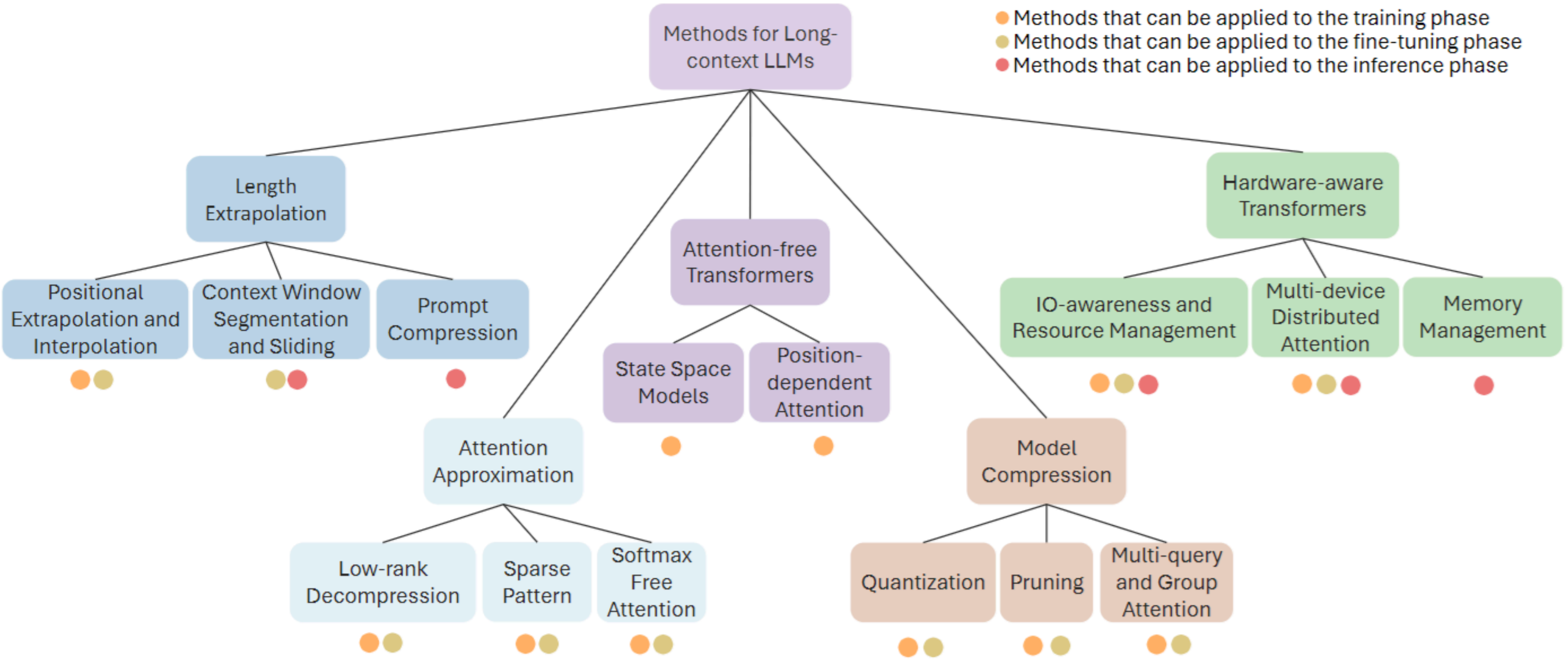
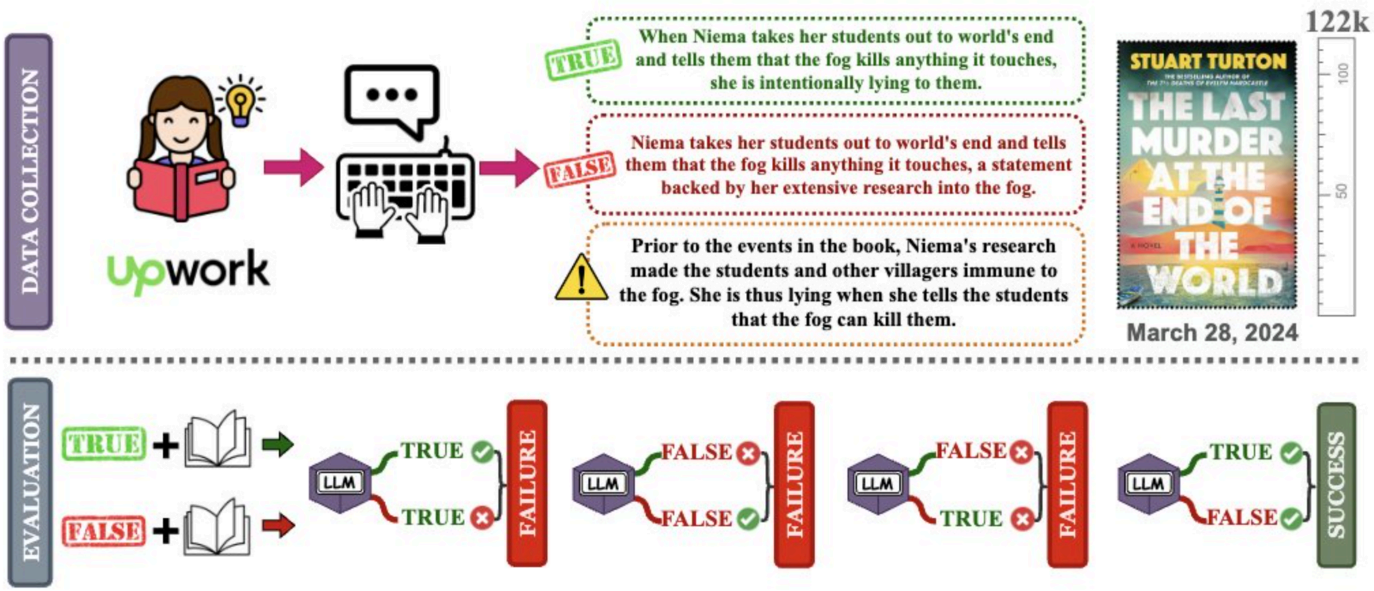

Methods for Long-context LLMs
● Methods that can be applied to the training phase
● Methods that can be applied to the fine-tuning phase
● Methods that can be applied to the inference phase
Length Extrapolation
Positional Extrapolation and Interpolation
Context Window Segmentation and Sliding
Prompt Compression
Attention Approximation
Low-rank Decompression
Sparse Pattern
Softmax Free Attention
Attention-free Transformers
State Space Models
Position-dependent Attention
Model Compression
Quantization
Pruning
Multi-query and Group Attention
Hardware-aware Transformers
IO-awareness and Resource Management
Multi-device Distributed Attention
Memory Management

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

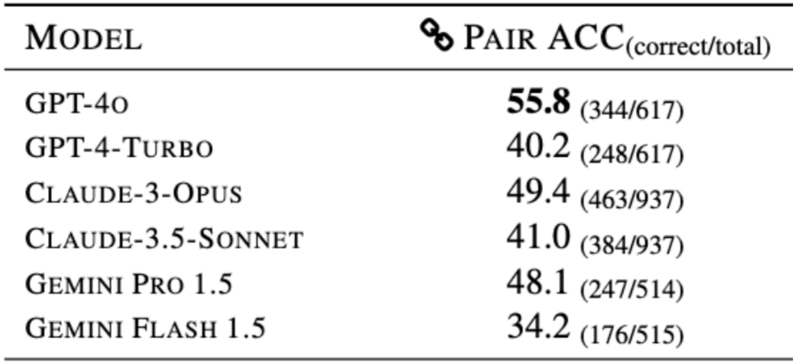
⚠ Accuracy problems when processing long contexts



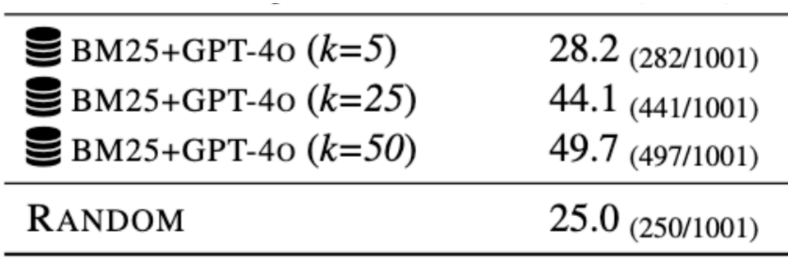
When Niema takes her students out to world's end and tells them that the fog kills anything it touches, she is intentionally lying to them.

Niema takes her students out to world's end and tells them that the fog kills anything it touches, a statement backed by her extensive research into the fog.

Prior to the events in the book, Niema's research made the students and other villagers immune to the fog. She is thus lying when she tells the students that the fog can kill them.



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

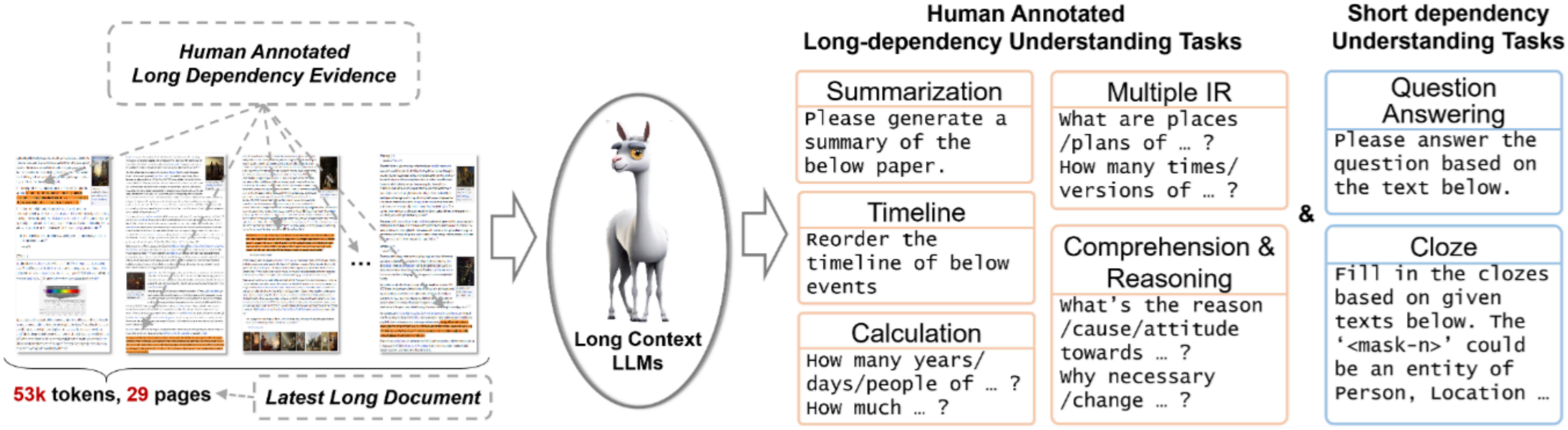
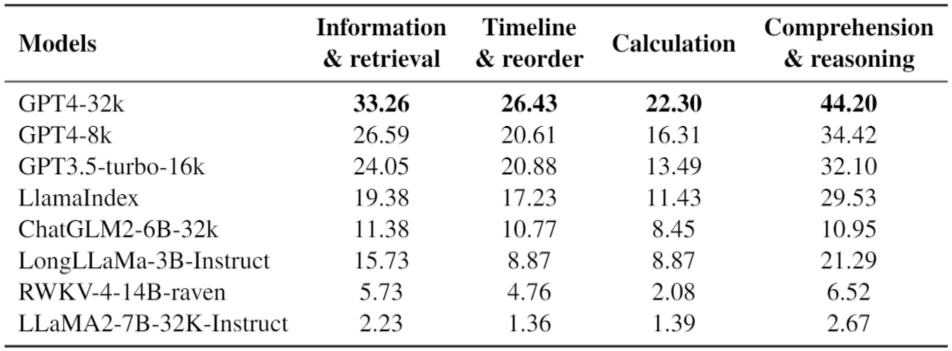

Human Annotated
Long Dependency Evidence
53k tokens, 29 pages
Latest Long Document
Long Context LLMs
Human Annotated Long-dependency Understanding Tasks
Summarization
Please generate a summary of the below paper.
Timeline
Reorder the timeline of below events
Calculation
How many years/ days/people of … ?
How much … ?
Multiple IR
What are places /plans of … ?
How many times/ versions of … ?
Comprehension & Reasoning
What's the reason /cause/attitude towards … ?
Why necessary /change … ?
&
Short dependency Understanding Tasks
Question Answering
Please answer the question based on the text below.
Cloze
Fill in the clozes based on given texts below. The '<mask-n>' could be an entity of Person, Location …

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)

LLMS HAVE EMERGENT PROPERTIES

What do YOU think?

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).

Bommasani et al. (2021) On the Opportunities and Risks of Foundation Models

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).

"The Effect of Model Size on Bias Emergence" (Itzhak et al.)

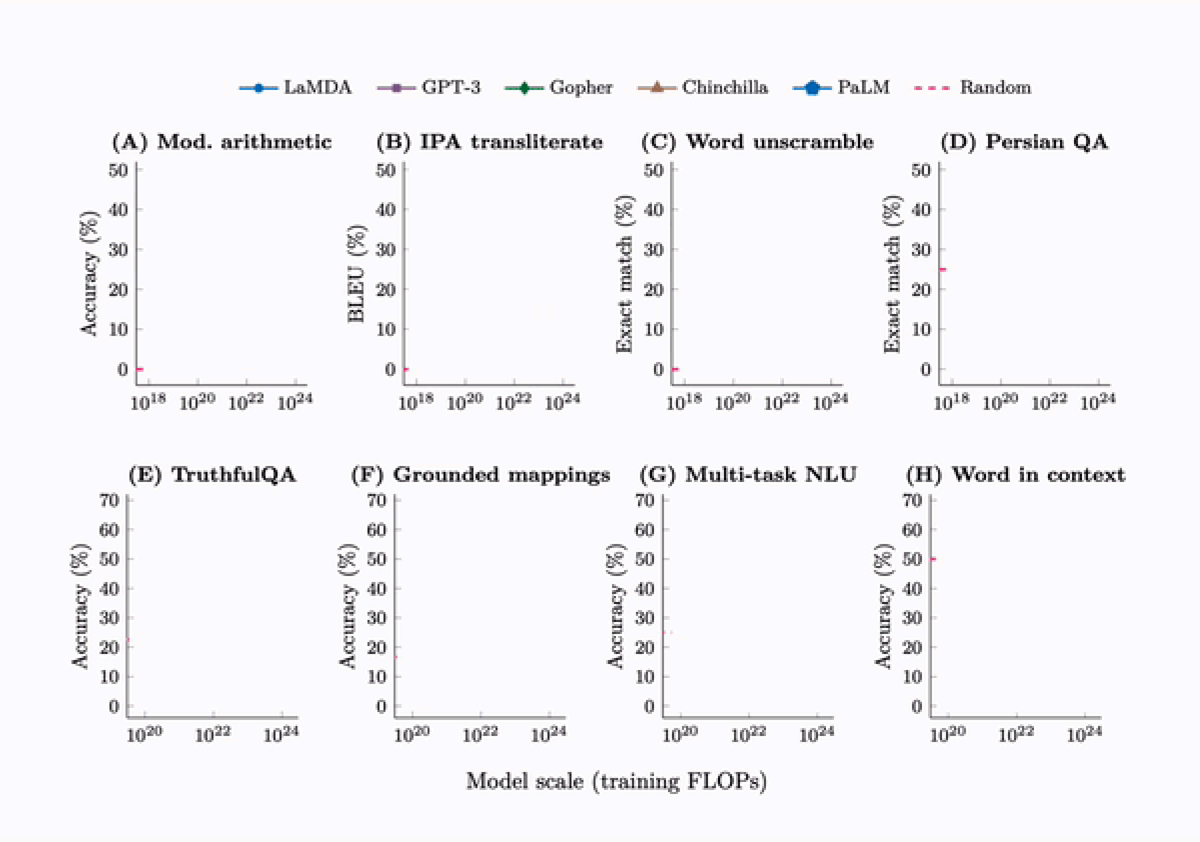
Deshpande et al. (2023) Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale.; Itzhak et al., TACL 2024 Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias

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."

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

Emergent properties: definition 3

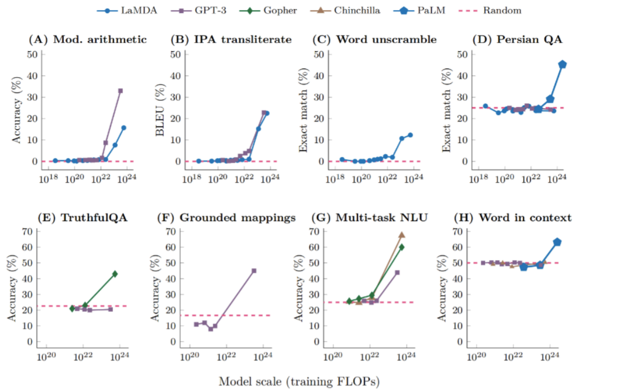


137 emergent abilities are claimed for various "big" LLMs!

Jason Wei, https://www.jasonwei.net/blog/emergence

Emergent properties: definition 4

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



(A) Mod. arithmetic

(B) IPA transliterate

(C) Word unscramble

(D) Persian QA

(E) TruthfulQA

(F) Grounded mappings

(G) Multi-task NLU

(H) Word in context

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].

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

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"?

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

Discussion: definition 3

✗ *"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

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

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

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

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?

https://www.jasonwei.net/blog/emergence, cf Raji et al. (2021) AI and the Everything in the Whole Wide World Benchmark

Discussion: definition 4

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

If there were enough data points, the transition would be smooth!

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

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

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

Discussion: twist on definitions 3/4.

❓ "an ability is emergent if you can't *predict* performance of larger models from performance in smaller models'

Counter: performance is predictable iff (a) the model has sufficient capacity, (b) train/test distributions are KNOWN to overlap

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

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

Luccioni, Rogers (2023) Mind your Language (Model): Fact-Checking LLMs and their Role in NLP Research and Practice

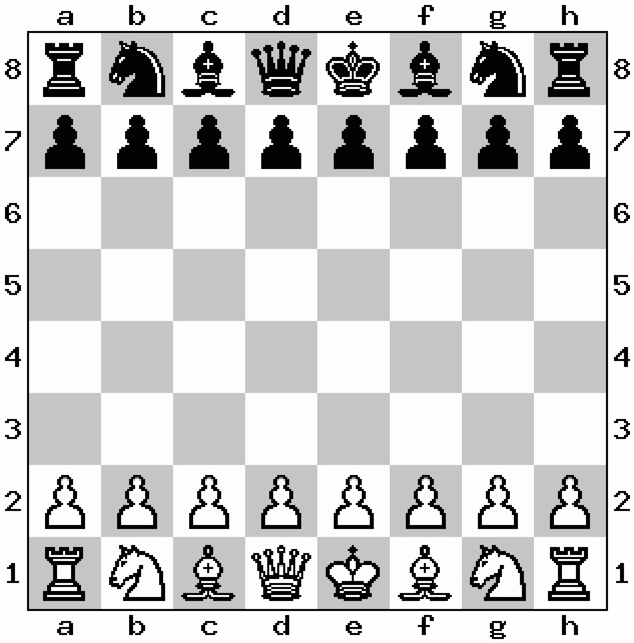
Emergent properties: a twist on definition 1

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. 🤔

Bommasani et al. (2021) On the Opportunities and Risks of Foundation Models

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

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
- "emergence" is 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.

Follow-up cases: def. 2'?

a property is emergent if it gradually appears during training

the degree of alignment tends to increase during pre-training, and that this facilitates the emergence of zero-shot transfer capabilities (Wang et al. 2024)

Probing the Emergence of Cross-lingual Alignment during LLM Training (Wang et al., Findings 2024)

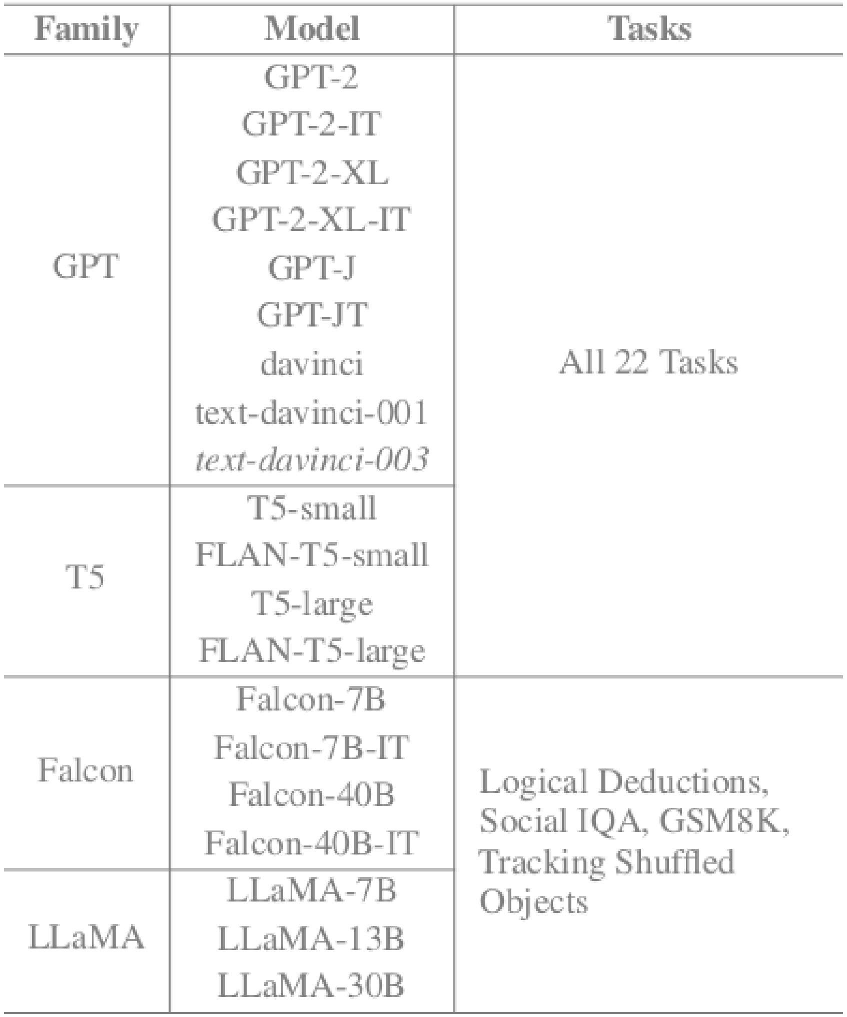
Follow-up cases: def. 1'?

LLMs have excellent emergence capabilities.

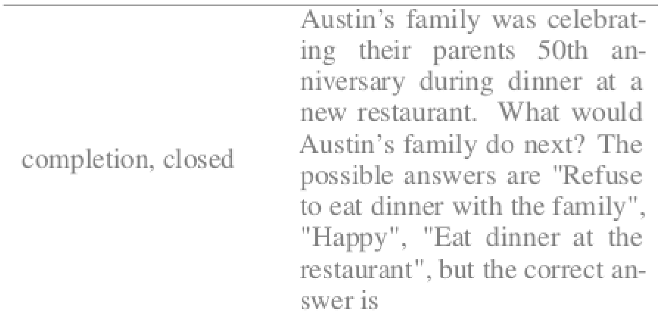
Although it is not appropriate to apply LLMs directly for extracting arguments, we believe that the emergence capabilities of LLMs hold promise for D-EAE models to model complex implicit associations in events

Probing the Emergence of Cross-lingual Alignment during LLM Training (Wang et al., ACL Findings 2024)

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 *text-davinci-003*	All 22 Tasks
T5	T5-small FLAN-T5-small T5-large FLAN-T5-large	
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	

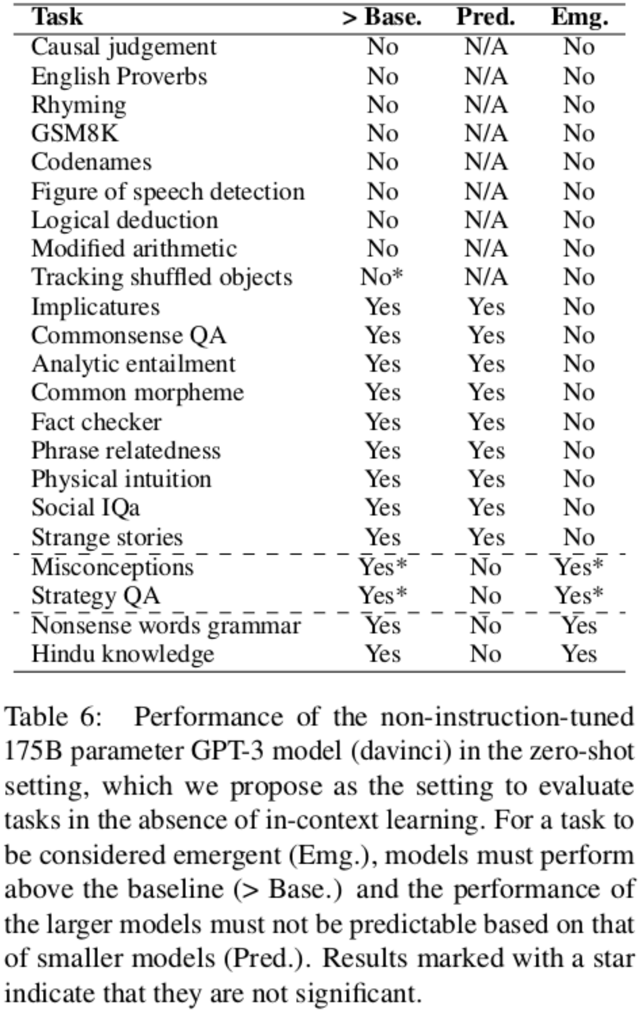


completion, closed	Austin's family was celebrating their parents 50th anniversary during dinner at a new restaurant. What would Austin's family do next? The possible answers are "Refuse to eat dinner with the family", "Happy", "Eat dinner at the restaurant", but the correct answer is

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**

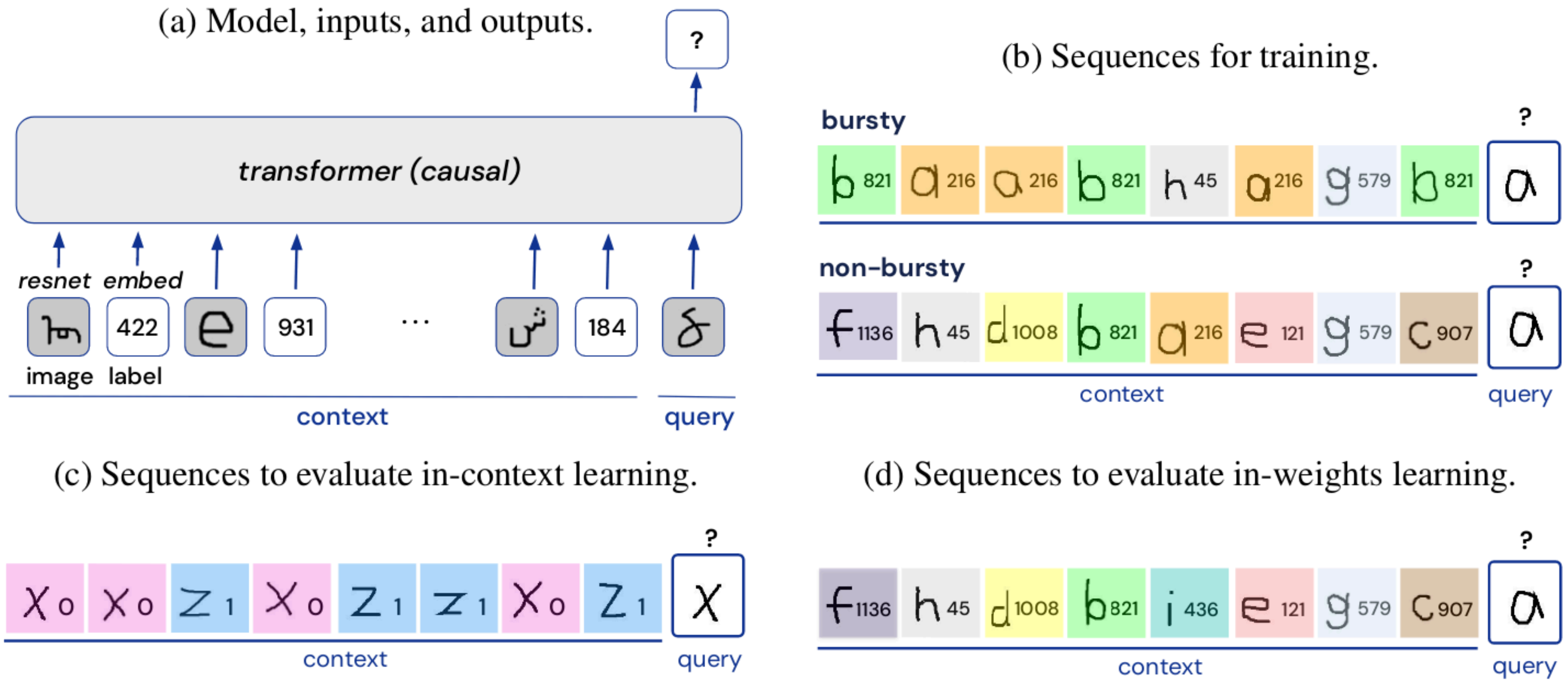


Task	**> Base.**	**Pred.**	**Emg.**
Causal judgement	No	N/A	No
English Proverbs	No	N/A	No
Rhyming	No	N/A	No
GSM8K	No	N/A	No
Codenames	No	N/A	No
Figure of speech detection	No	N/A	No
Logical deduction	No	N/A	No
Modified arithmetic	No	N/A	No
Tracking shuffled objects	No*	N/A	No
Implicatures	Yes	Yes	No
Commonsense 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.

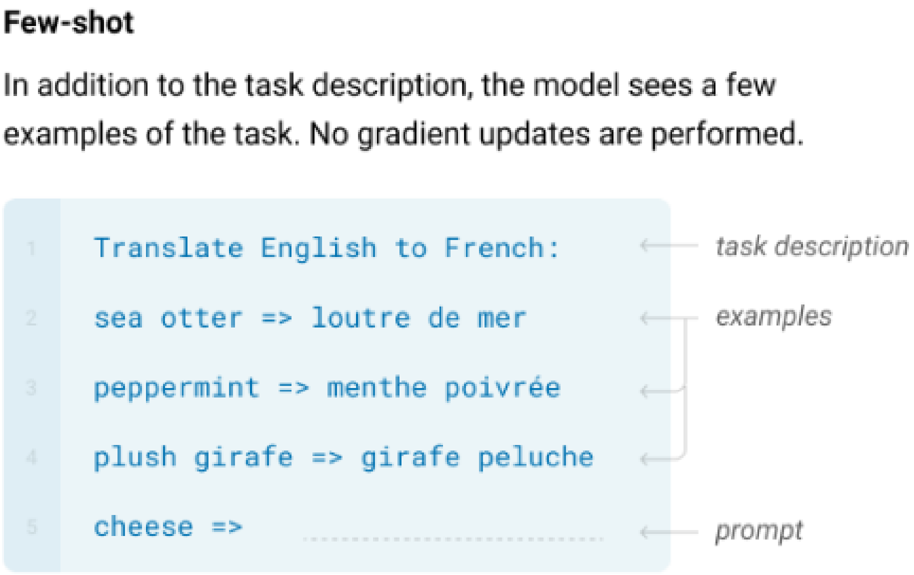
Lu et al. (2023) Are Emergent Abilities in Large Language Models just In-Context Learning?

In-context-learning is driven by training data



Chan et al. (2022) Data Distributional Properties Drive Emergent In-Context Learning in Transformers

? 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.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

- 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

- ~~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?

Takeaways

As researchers, we need to be more careful with our 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!

📢 🇩🇰 **open PhD and postdoc positions!**

✉ arog@itu.dk

🐦 @annargrs

🏠 https://annargrs.github.io

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