Behind the Secrets of Large Language Models

Lecture 9

Evaluating LLMs

Simon Razniewski

ScaDS.AI
DRESDEN LEIPZIG

TECHNISCHE UNIVERSITÄT DRESDEN

Nov 22, 2024

TUD'S CAPELLA SUPERCOMPUTER RANKS THIRD MOST POWERFUL SUPERCOMPUTER IN GERMANY AND IS FIFTH WORLDWIDE FOR ENERGY EFFICIENCY

With more than 38 Petaflops/s of high-performance computing power (number of double-precision floating-point operations performed per second), ZIH's Capella system achieved outstanding results in November's edition of the Top500 list of the most powerful supercomputers. Installed by Megware, a company based in Saxony, and designed in close collaboration with Lenovo and others, this high-performance computer is equipped with over 140 nodes. Every node has four NVIDIA H100 accelerators and two AMD processors, each with 32 cores. In addition, each graphics processing unit (GPU) is equipped with 94 Gigabytes of high-bandwidth memory. A high-speed storage system with over 1 petabyte is used as a burst buffer and provides data to the AI accelerators at over 1500 gigabytes per second, benefiting data-intensive applications such as training large AI models. By integrating Capella into the data center's existing infrastructure, the ZIH was also able to efficiently connect the file systems of the Barnard HPC cluster in the overall complex. With its hot water cooling and use of waste heat, Capella upholds the high energy efficiency standards set by the ZIH data center.

Recap: Lecture so far

- L1-4: Neural networks, deep learning, word embeddings
- L5: Obtain training data
- L6, L8: Choose architecture
- L7: Train your model
  - Requires an evaluation metric, typically, loss
  - But loss only measures performance on training objective

- **Today: Holistic look at evaluating LLMs**

5

Outline

1. Evaluation generics
    1. Hardness
    2. Stages and metrics
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

Acknowledgment/License

- Shared under CC BY 4.0

Outline

1. **Evaluation generics**
    1. **Hardness**
    2. **Stages and metrics**
    3. **Statistics 101**
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

Why do we need evaluation?

- Show off? ☺

> Llama 3.3 is a text-only 70B instruction-tuned model that provides enhanced performance relative to Llama 3.1 70B–and to Llama 3.2 90B when used for text-only applications. Moreover, for some applications, Llama 3.3 70B approaches the performance of Llama 3.1 405B.

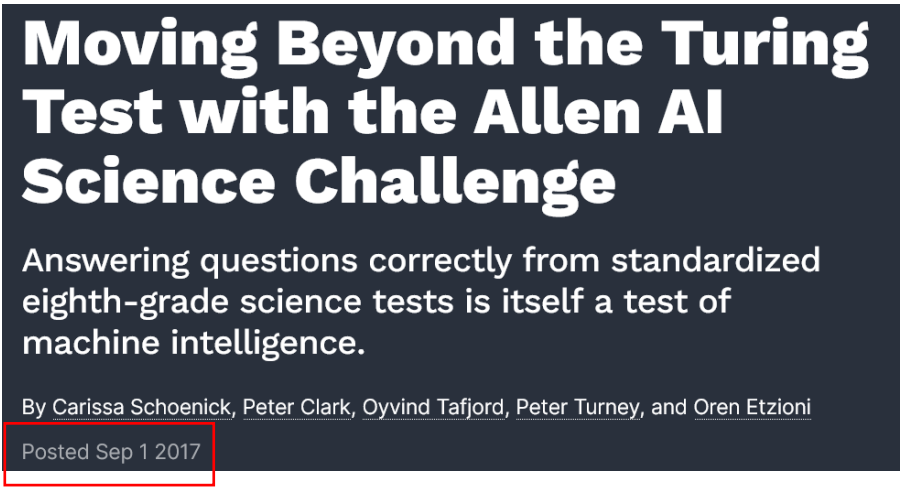
- Static view: Economic efficiency!
  - Stiftung Warentest/online shopping reviews

- Dynamic view: Measure progress!
  - *“What you can’t measure you cannot improve”*

Outline

1. **Evaluation generics**
    1. **Hardness**
    2. Stages and metrics
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

Problem 1: Benchmarks saturation/ short significance



Problem 2: Leakage

NLP Evaluation in trouble:
On the Need to Measure LLM Data Contamination for each Benchmark

Oscar Sainz[1] **Jon Ander Campos**[2] **Iker García-Ferrero**[1] **Julen Etxaniz**[1]
Oier Lopez de Lacalle[1] **Eneko Agirre**[1]

[1] HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{oscar.sainz,iker.graciaf,julen.etxaniz}@ehu.eus
{oier.lopezdelacalle,e.agirre}@ehu.eus

LAST UPDATED SEPTEMBER 6, 2023 IN AI ORIGINS & EVOLUTION

The Problems with LLM Benchmarks

AI benchmarks are flawed - with dataset contamination, biases and are often not representative of real world use cases. But what are the alternatives?

OpenAI didn't release much information about GPT-4 — not even the size of the model — but heavily emphasized its performance on professional licensing exams and other standardized tests. For instance, GPT-4 reportedly scored in the 90th percentile on the bar exam. So there's been much speculation about what this means for professionals such as lawyers.

We don't know the answer, but we hope to inject some reality into the conversation. OpenAI may have violated the cardinal rule of machine learning don't test on your training data. Setting that aside, there's a bigger problem.

***Stop Uploading Test Data in Plain Text*: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks**

Alon Jacovi[1] **Avi Caciularu**[1,2] **Omer Goldman**[1] **Yoav Goldberg**[1,3]

[1] Bar Ilan University
[2] Google Research
[3] Allen Institute for Artificial Intelligence

Pseudo solution: Private benchmarks

- Benchmark host does not publish the test data
- Requires contestants to submit their models, run and evaluate them on the benchmark's servers
- But:
  - Requires complex compute setup on side of benchmark host
  - Incentives in academia do not favor this
  - For commercial secrets like GPT-4 unlikely that developers would hand them out

Problem 3: Evaluation format

- True/false evaluations only possible for tasks that have one or a handful of correct answers
  - Generally only the case in a few domains, e.g., Math
  - Not for general QA, translation, summarization, etc.
- Open-ended tasks generally much harder to evaluate
  - Mirrors a dilemma that also (university) teachers face
- Common approach:
  - Where possible, transform other tasks into multiple-choice form
  - Restrict evaluation to multiple-choice tasks

Problem 4: Designing bias-free benchmarks is hard

VQA Jabri et al, 2017

"Overall, our results suggest that the performance of current VQA systems is not significantly better than that of systems designed to exploit dataset biases."

Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment

Masatoshi Tsuchiya
Toyohashi University of Technology
1–1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, Japan
tsuchiya@imc.tut.ac.jp

Premise: A cat is sleeping on the sofa.
Hypothesis: The cat is resting.

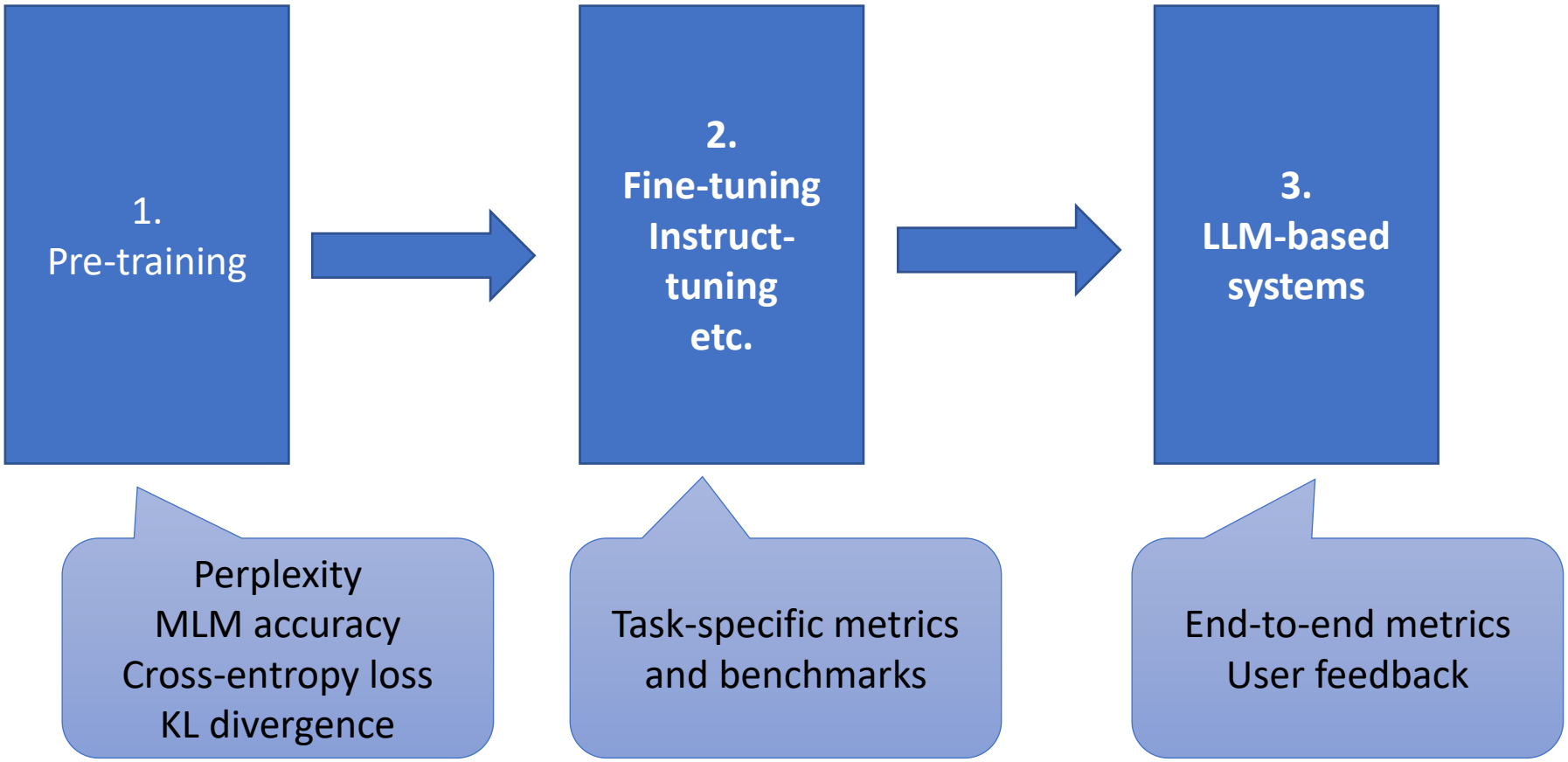
Premise: A man is eating an apple.
Hypothesis: A man is eating a banana.

Predicted labels	Corpus labels		
	Entailment	Neutral	Contradiction
Entailment	2275	644	706
Neutral	508	1976	563
Contradiction	585	599	1968

Outline

1. **Evaluation generics**
    1. Hardness
    2. **Stages and metrics**
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

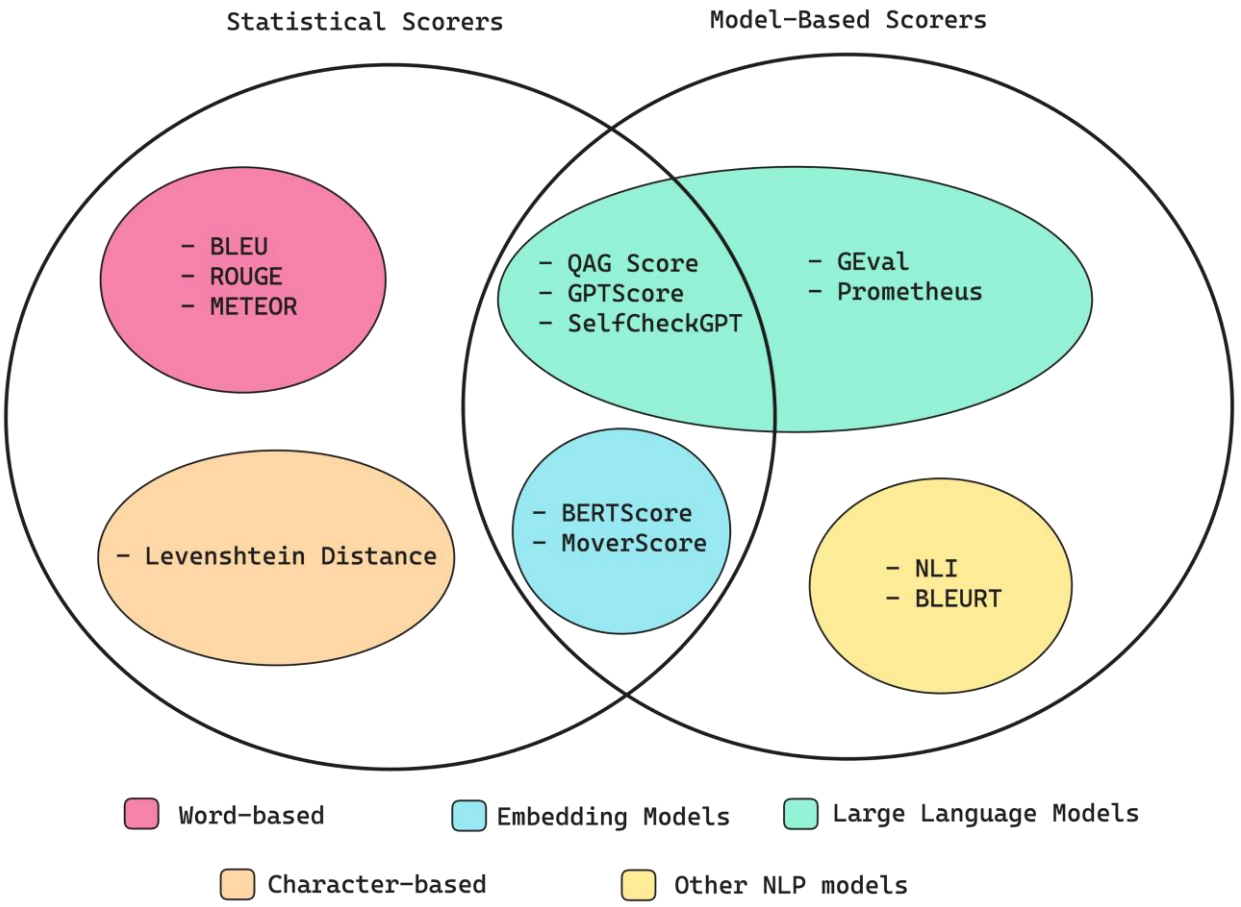
What stage to evaluate?



Metrics for fine-tuned models

- Multiple-choice/classification easiest to evaluate
  - Accuracy
  - P(recision)
  - R(ecall)
  - F1
  - P@1
  - R@Pxy
  - …
- Open-ended tasks:
  - Text similarity metrics
  - (un-)supervised models

Open-ended task evaluation



https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation

Open-ended generation

- Metrics strongly motivated from machine translation field
  - There are often several acceptable translations
  - *"She entered the room" – "Sie betrat das Zimmer", "Sie trat in das Zimmer ein", "Sie hat den Raum betreten"*

- Levenshtein distance: Number of character replace/add/remove operations between two strings
  - *Lev(Sie betrat das Zimmer, Sie trat in das Zimmer) =* 4
  - *Lev(Sie betrat das Zimmer, Sie trat in das Zimmer ein) =* 9
  - *Lev(Sie betrat das Zimmer, Sie hat den Raum betreten) =* 16
  - *Lev(Sie betrat das Zimmer, Er aß Tomaten mit Ingwer) =* 18

https://planetcalc.com/1721/

Open-ended generation (2)

- Beyond characters:
  - BLEU (bilingual evaluation understudy): word overlap
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation): word overlap, recall-oriented

- Syntactic overlap metrics still heavily limited
  - *I love going for a walk early in the morning.*
  - *At the start of the day, I enjoy taking a stroll.*

→Model-based metrics like BERTScore measure embedding similarity
  - Performance-interpretability tradeoff

Open-ended generation (3)

- *"Generate a 5-page patent application"*
  - At scale, also embedding metrics get blurred
  - Long-document evaluation unsolved

Dataset Split	**test**					
Metric	BERTScore			ROUGE-L		
Outline	short	medium	long	short	medium	long
Llama-3 8B	67.8	68.3	68.7	23.7	23.0	23.9
Llama-3 70B	68.6	69.5	70.2	24.7	23.2	24.8
Mixtral-8x7B	68.1	68.3	69.1	23.6	23.1	25.4
Qwen2-72B	69.0	69.4	70.2	25.7	24.9	26.1
Llama-3 8B SFT	**69.2**	**69.9**	**70.5**	**26.0**	**27.2**	**26.5**

[Knappich et al., Arxiv 2024]

Metrics for LLM-based systems

- Task performance matters
- But tasks and uses cannot entirely be foreseen

- Beyond task performance: General human alignment
  - Truthfulness
  - Safety
  - Bias and fairness
  - Robustness and security
  - Privacy, unlearning, and copyright implications
  - Calibration and confidence
  - Transparency and causal interventions

→ Human labels/preference

Outline

1. **Evaluation generics**
    1. Hardness
    2. Stages and metrics
    3. **Statistics 101**
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

Evaluate how much?

- Benchmarks typically are of moderate size (100s-1000s of records)
- But there are millions or more of possible tasks instances
- Are 250 records enough to decide which model is better?

Statistical significance

- Are 250 records enough?

- It depends!
  - To prove at 95% confidence that a system with 83% sample accuracy is better than one with 43% yes
  - To prove that at 95% confidence that a system with 83% sample accuracy is better than one with 81% probably not

→ Law of large numbers, statistical significance tests

Tests for statistical significance

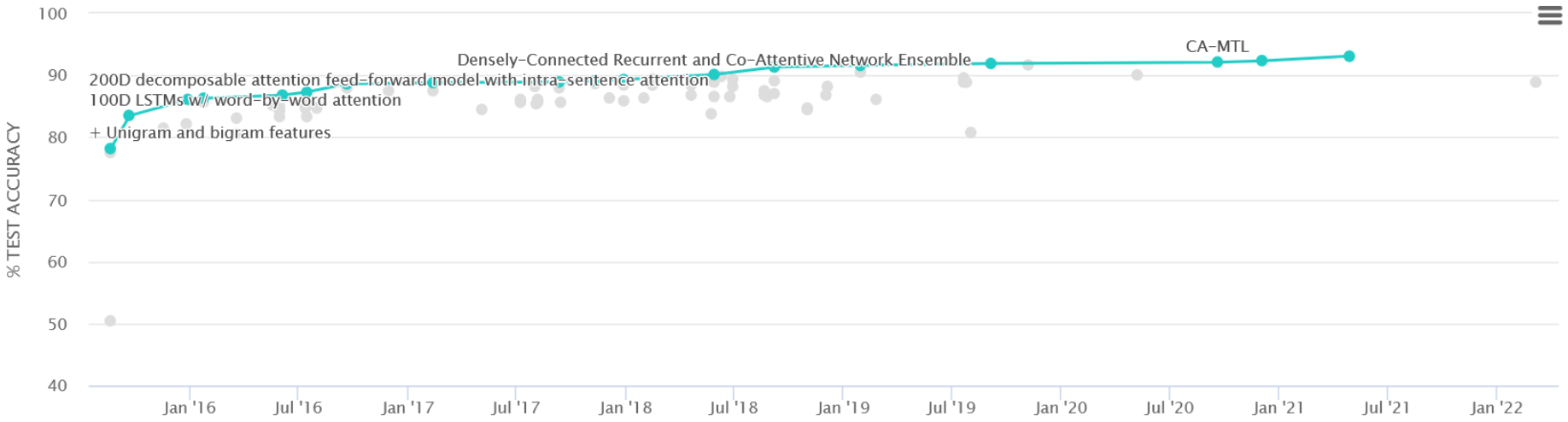
- t-test, z-test, McNemar's test, exact binomial test, …
- Actually all make assumptions about source distribution (or independence)
- Safe choice:
  - Paired bootstrap resampling
    - Subsample your test set 1000 times, see whether system A beats system B in >95% of cases

Code!

[Statistical Significance Tests for Machine Translation Evaluation, The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing]

Side note: When is a dataset solved?

- Ground truth virtually always contains errors
- Some ambiguity virtually always hidden in tasks
  - Leaderboards rarely go beyond 90-95%



Natural Language Inference on SNLI

Outline

1. Evaluation generics
    1. Hardness
    2. Stages and metrics
    3. Statistics 101
2. **Important benchmarks**
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

Benchmark Desiderata

- Superhuman scaling
  - Doesn’t saturate quickly, can scale beyond human-level
- Ease of setting up
  - Does not require specialized training or complicated software (e.g., no 1 TB text corpus in the backend needed)
- Stable
  - Does not represent highly volatile knowledge that penalizes up-to-date models (e.g., “who is the current US president?”)
- Automatic evaluability
  - Fast feedback loops means no humans are allowed

Benchmark Desiderata (2)

- Clear downstream implications
  - ↑benchmark → ↑downstream tasks,
    or ↑benchmark → new methods that ↑downstream
- The metric is interpretable
  - Accuracy is more interpretable than nats/bits
- Useful for hill climbing on
  - has progression, performance not an indicator function but smooth

MMLU: Measuring Massive Multitask Accuracy

- Massive Multitask Language Understanding
- Includes questions on topics ranging from mathematics and physics to history and law. Difficulty ranges from highschool to professional exams.
- ~15k questions

Find all c in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.
(A) 0 **(B) 1** (C) 2 (D) 3

Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?
(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders.
(B) Yes, if Hermit was responsible for the explosive charge under the driveway.
(C) No, because Seller ignored the sign, which warned him against proceeding further.
(D) No, if Hermit reasonably feared that intruders would come and harm him or his family.

MMLU - Scope

57 task types

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s²
(B) more than 9.8 m/s²
(C) less than 9.8 m/s²
(D) Cannot say unless the speed of throw is given.

College Mathematics

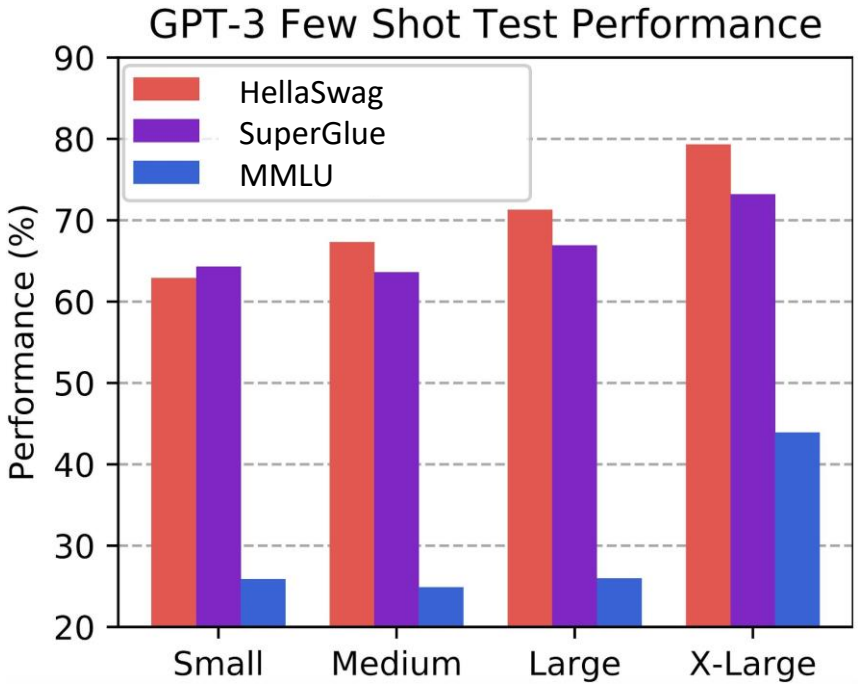
In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points
(B) circle
(C) half-line
(D) line

Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?
(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders.
(B) Yes, if Hermit was responsible for the explosive charge under the driveway.
(C) No, because Seller ignored the sign, which warned him against proceeding further.
(D) No, if Hermit reasonably feared that intruders would come and harm him or his family.

MMLU - Capabilities

- Results of GPT-3 (SOTA as of benchmark release)

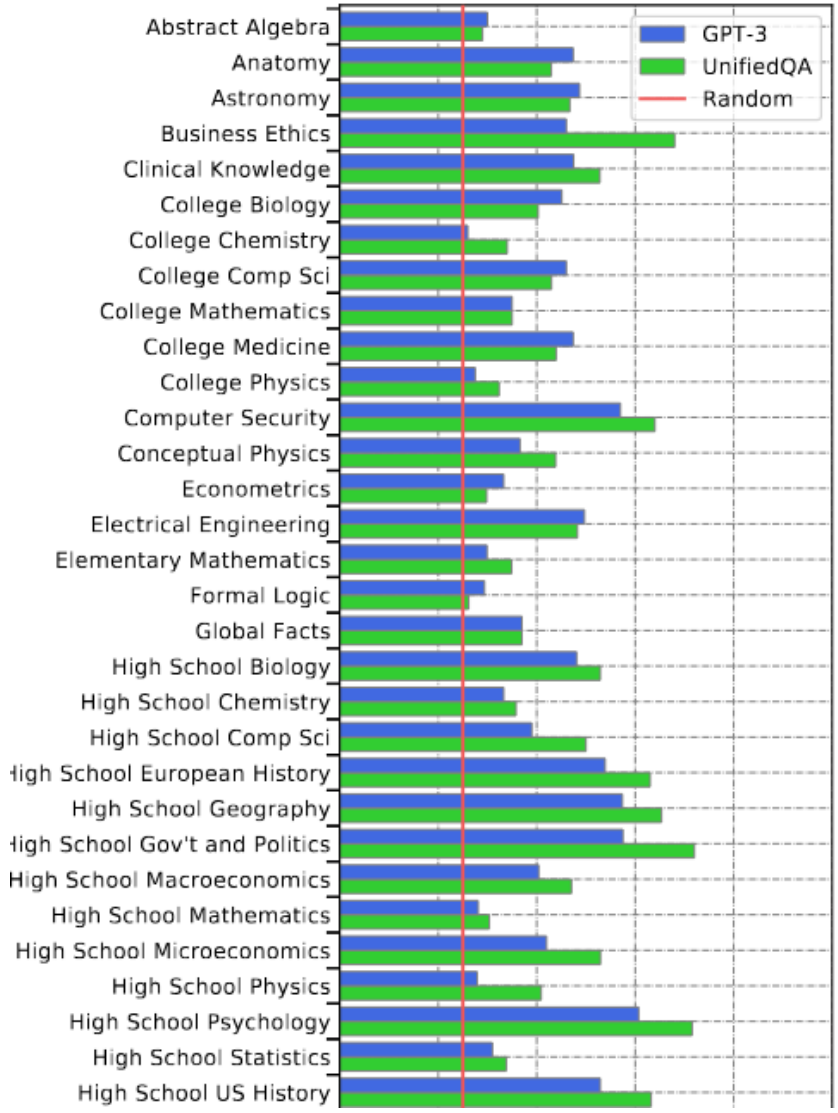


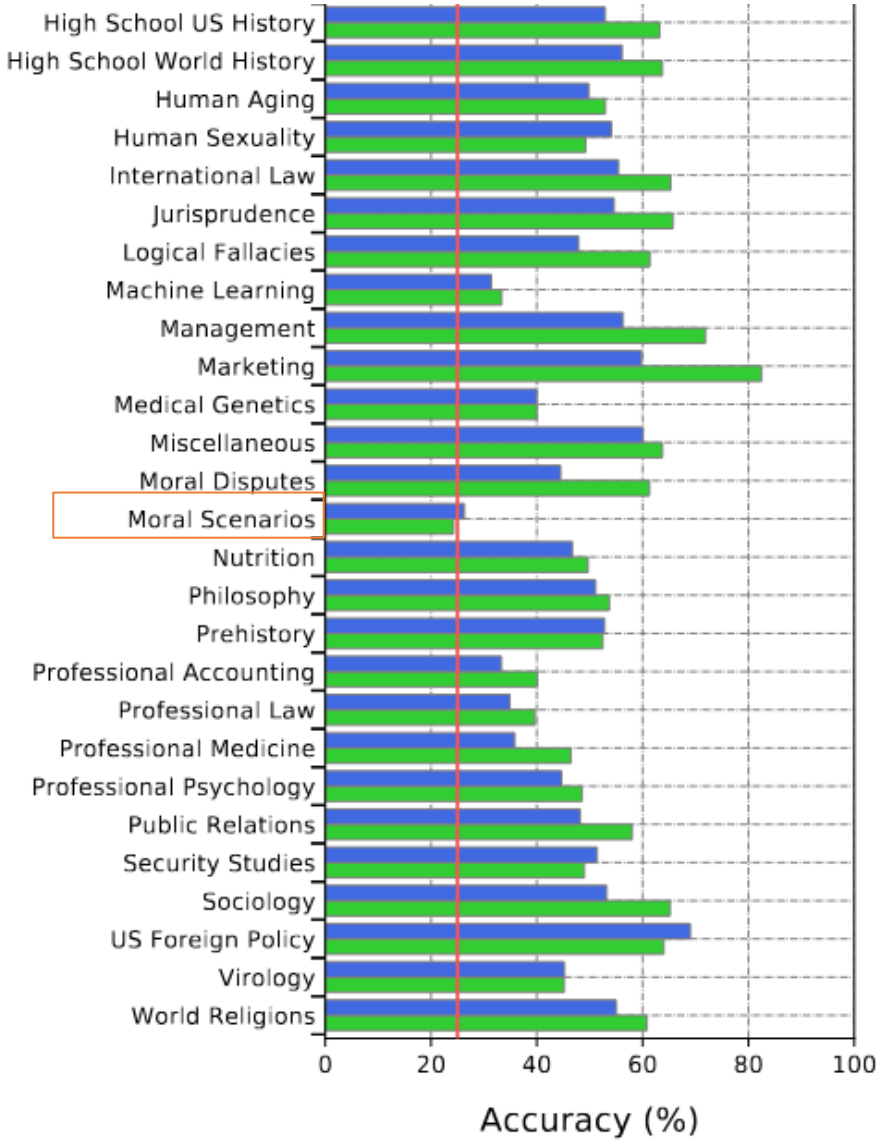
Baseline: 25%

Accuracy breakdown by broad subject area

Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
UnifiedQA	45.6	56.6	40.2	54.6	48.9
GPT-3 Small	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium	26.1	21.6	25.6	25.5	24.9
GPT-3 Large	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large	40.8	50.4	36.7	48.8	43.9
Layperson (human)					34.5
Expert (human)					89.8

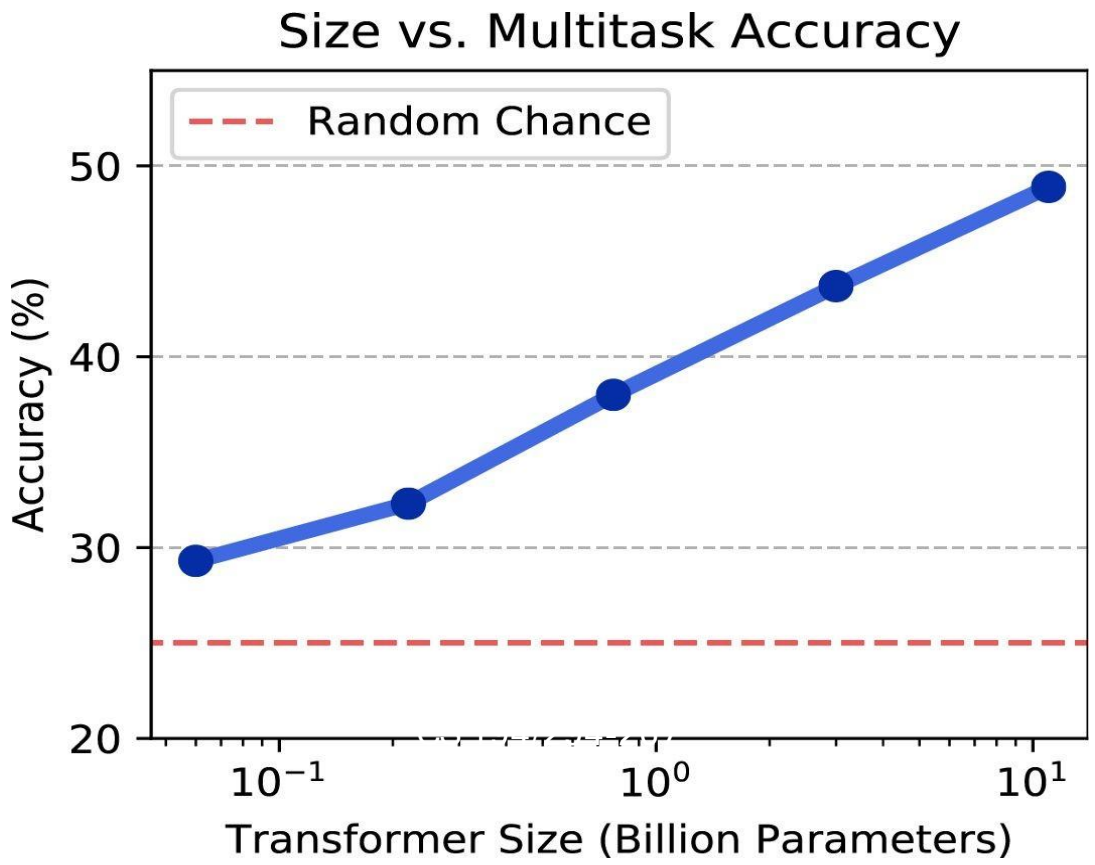
Per-Subject Breakdown



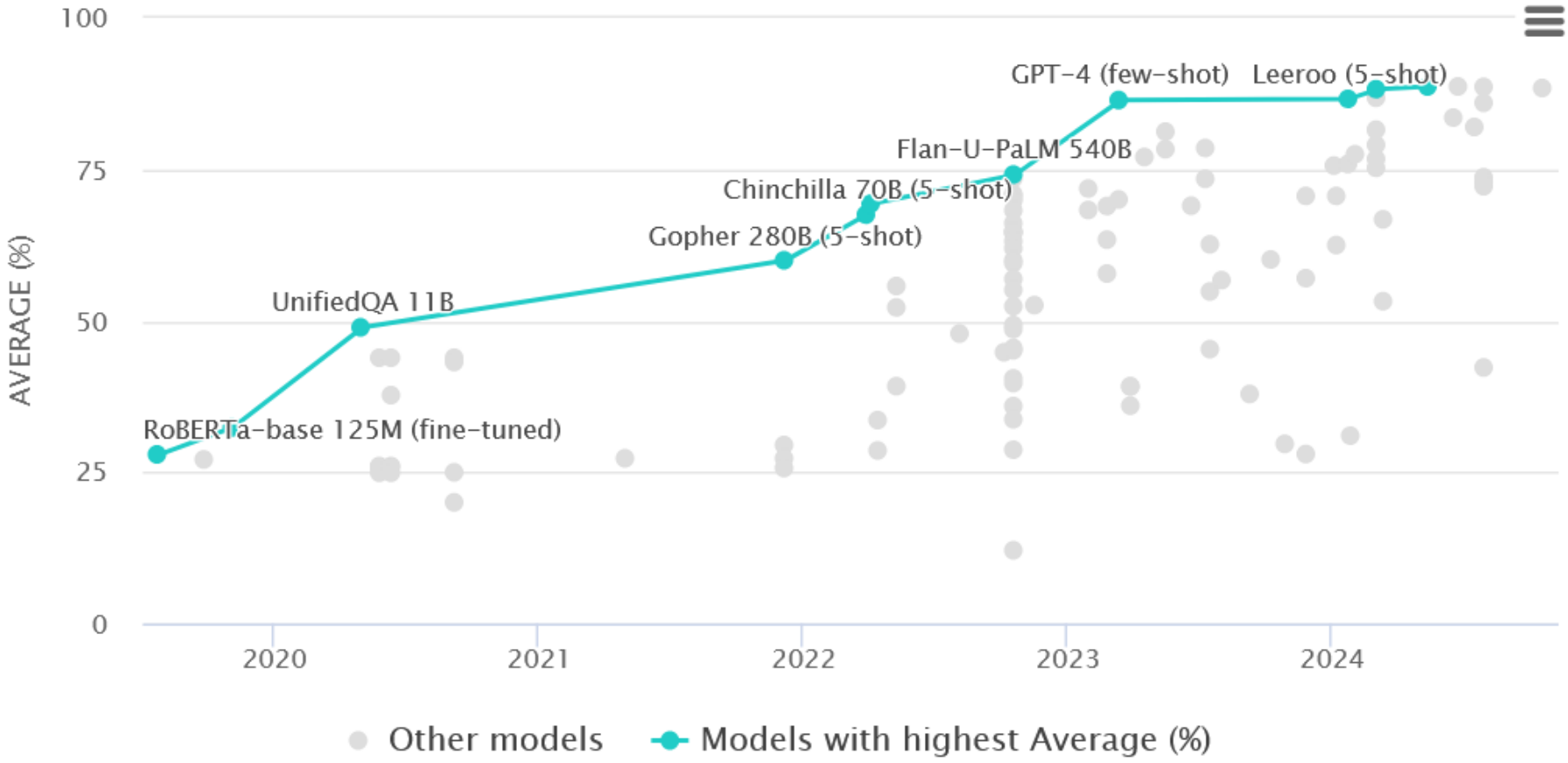


Scaling on MMLU

~100 trillion parameters leads to an 85% average?



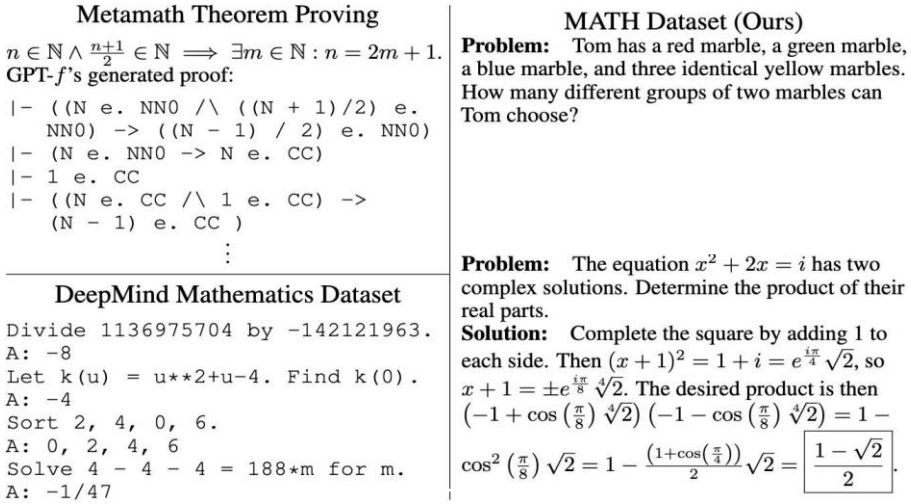
The reality in 2024



https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

MATH - Introduction

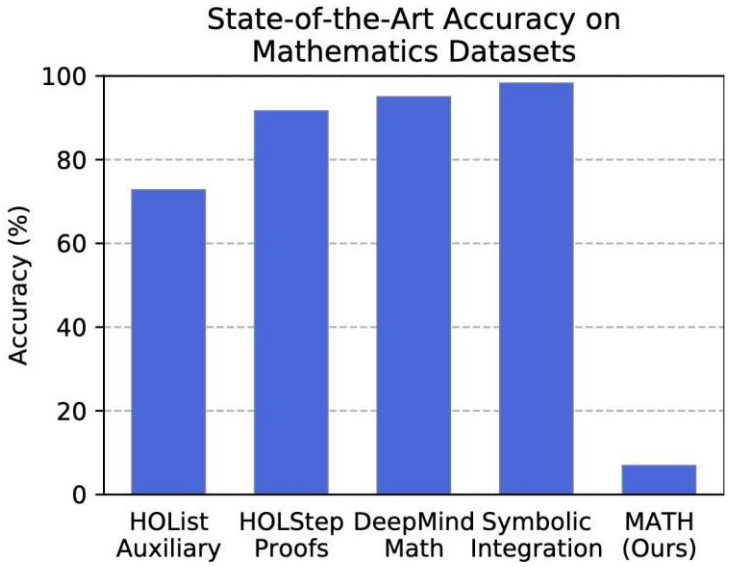
- Dataset of 12,500 challenging competition mathematics problems
- Each problem has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations.



https://paperswithcode.com/dataset/math

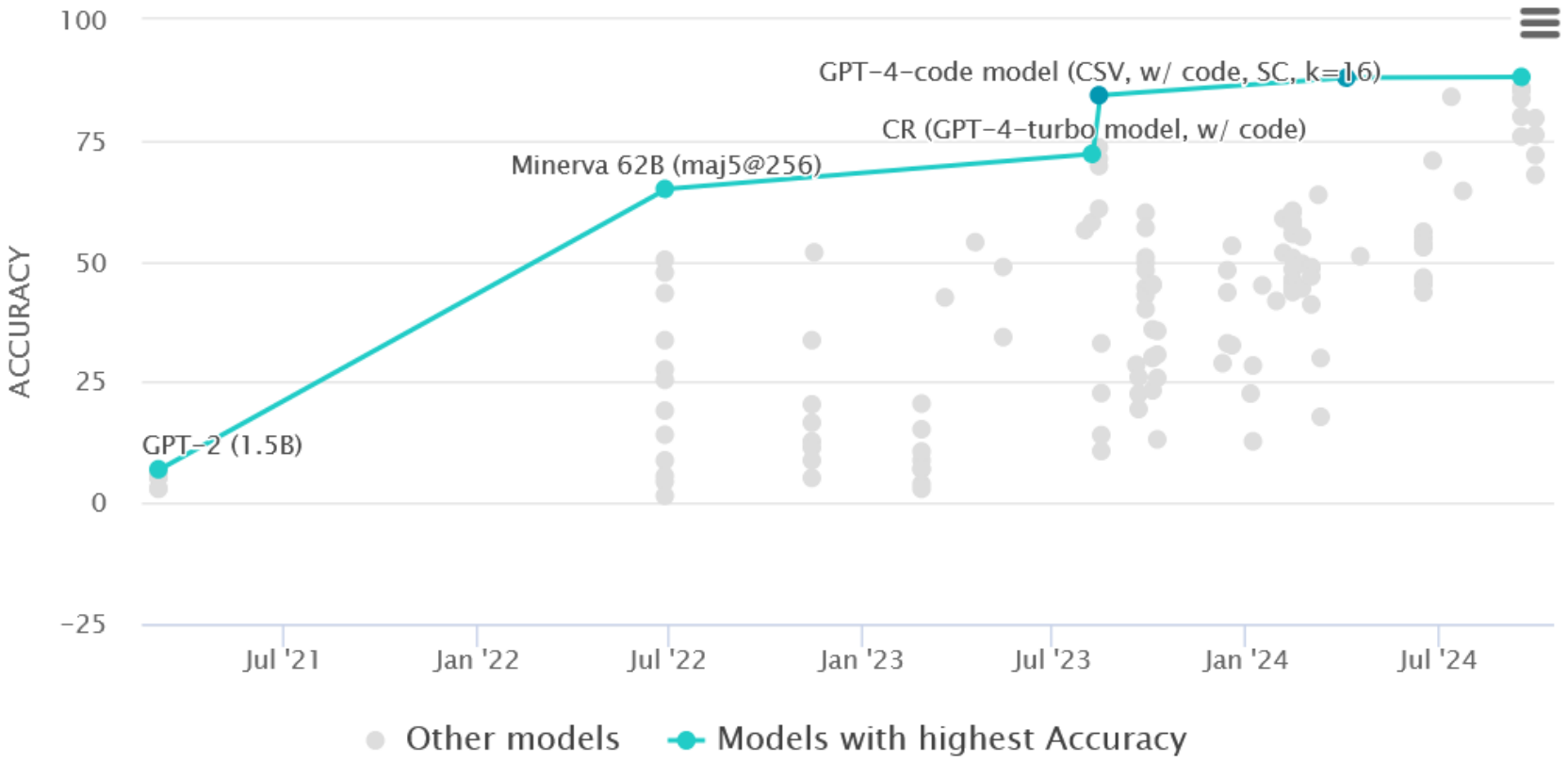
MATH - Difficulty

At its release, MATH was extremely challenging: large language models achieved accuracies ranging from 3.0% to 6.9%.



MATH - Increases in Capabilities

Recent capabilities improvements have demonstrated a significant increase in performance, even with small models.



https://paperswithcode.com/sota/math-word-problem-solving-on-math

LMArena

- Humans label LLM answers to arbitrary prompts by binary preference
  →Elo ranking like in chess

- Upside:
  - No restriction to domain
  - Human opinion is what often matters most
- Downside:
  - New models need time to enter ranking

https://lmarena.ai/

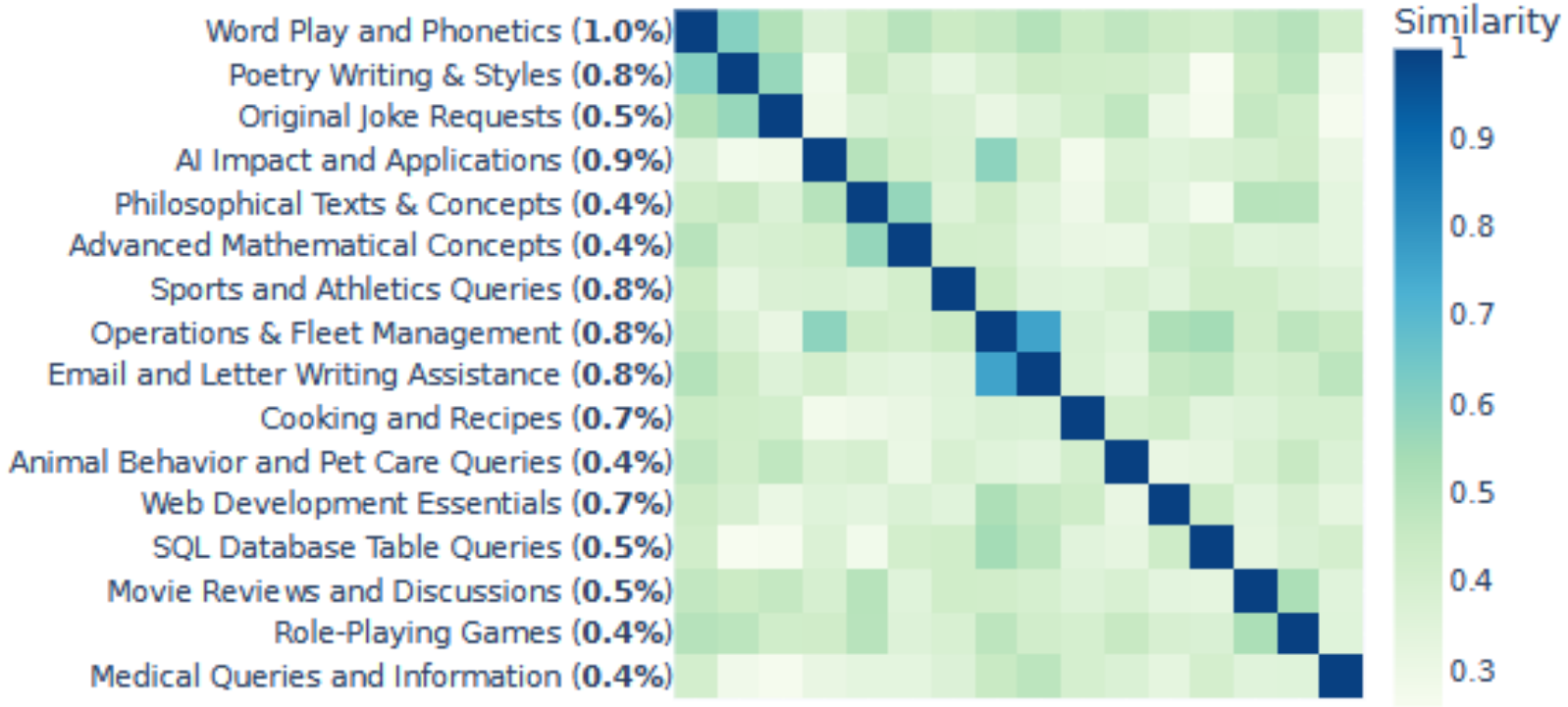


Figure 3. Similarity matrix of top-16 topic clusters. The number followed by the topic label represents the cluster size in percentage. Note that similarity is computed by cluster's centroid embeddings, hence diagonals are always one.

Outline

1. Evaluation generics
    1. Hardness
    2. Stages and metrics
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. **LLM-as-judge**
4. Case study: Evaluating factual knowledge
5. Discussion
    - Group work

LLM-as-judge: Motivation

- Existing automated benchmarks
  - Many LLM tasks not suited for multiple-choice format
  - BLEU, BERTScore, etc. do not capture “deep” semantics
  - Cannot keep up with moving LLM target

- Manual annotation of LLM outputs
  - Monetarily and time-wise difficult
    - Repeated runs during development infeasible
    - Ethical concerns, IRBs
  - Experts often hard to find
    - E.g., patent domain

- LLMs are automated systems with significant semantic knowledge
  → Can we use them for evaluating LLMs?

[Judging llm-as-a-judge with mt-bench and chatbot arena]

LLM-as-a-Judge

Question:

Write a travel blog post about Hawaii.

Model A:

Title: Discovering the Aloha Spirit: A Recent Trip to Hawaii Introduction: Hawaii, the 50th state of the United States ...

Model B:

Title: Aloha, Hawaii! A Cultural and Natural Paradise Awaits Subheading: Uncovering the Rich Culture and ...

Which model's response is better?

LLM Judge: GPT-4 / GPT-3.5 / Claude

Judgement:

I think A provides .. Therefore, A is better.

✅ Scalable
✅ Explainable

Specific LLM-judge-settings

- **Single Output Scoring w/o reference**
  - Judge LLM is provided with a scoring rubric as the criteria
  - Prompted to assign a score to LLM responses
  - Chain-of-thought helps: Verbal answer first, then score
- **Single Output Scoring w/ reference**
  - As above, but with reference answer
  - Helps to calibrate the scores
- **Pairwise Comparison**
  - Given two LLM generated outputs, the judge LLM will pick which one is the better generation with respect to the input
  - Often helpful when overall quality is high (e.g., coherence)

Biases and limitations of LLM-as-a-Judge

- **Position bias**
  - LLMs favor the answer in the first position.

  Verbosity bias
  - LLMs favor longer answers.

- **Self-appreciation bias**
  - LLMs favor its own answer or answers similar to its own answer

- **Limited reasoning ability**
  - LLMs fail to judge hard math/reasoning/code questions

Solutions and Results

- **Solutions:**
  - Swapping positions
  - Reference-guided judge
  - Chain-of-thought judge
  - Fine-tuned local judge
  - …

- **Results:**
  - The agreement between GPT-4 judges and humans reaches over 80%, the same level of agreement among humans.

Example evaluation prompt

https://chatgpt.com/
https://platform.openai.com/docs/overview

Explain the moon landing to a 5-year old, in one paragraph.

Explain the moon landing to a 12-year old, in one paragraph.

Evaluate how appropriate the following text is for a 6-year old.
Use three criteria: 1) Vocabulary 2) Grammatical complexity 3) Age-appropriate explanation.

Assign a score from 1 (worst) to 10 (best) to each criterion.

LLMs instead of Human Judges?
A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco[1], Raffaella Bernardi[2], Leonardo Bertolazzi[2], Desmond Elliott[3], Raquel Fernández[1], Albert Gatt[4], Esam Ghaleb[1], Mario Giulianelli [5], Michael Hanna[1], Alexander Koller[6], André F. T. Martins[7], Philipp Mondorf[8], Vera Neplenbroek[1], Sandro Pezzelle[1], Barbara Plank[8], David Schlangen[9], Alessandro Suglia[10], Aditya K Surikuchi[1], Ece Takmaz[4], Alberto Testoni[1]

[1]University of Amsterdam, [2]University of Trento, [3]University of Copenhagen, [4]Utrecht University, [5]ETH Zürich, [6]Saarland University, [7]Universidade de Lisboa & Unbabel, [8]LMU Munich & MCML, [9]University of Potsdam, [10]Heriot-Watt University,

Abstract

There is an increasing trend towards evaluating NLP models with LLM-generated judgments instead of human judgments. In the absence of a comparison against human data, this raises concerns about the validity of these evaluations; in case they are conducted with proprietary models, this also raises concerns over reproducibility. We provide JUDGE-BENCH, a collection of 20 NLP datasets with human annotations, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show that each LLM exhibits a large variance across datasets in its correlation to human judgments. We conclude that LLMs are not yet ready to systematically replace human judges in NLP.

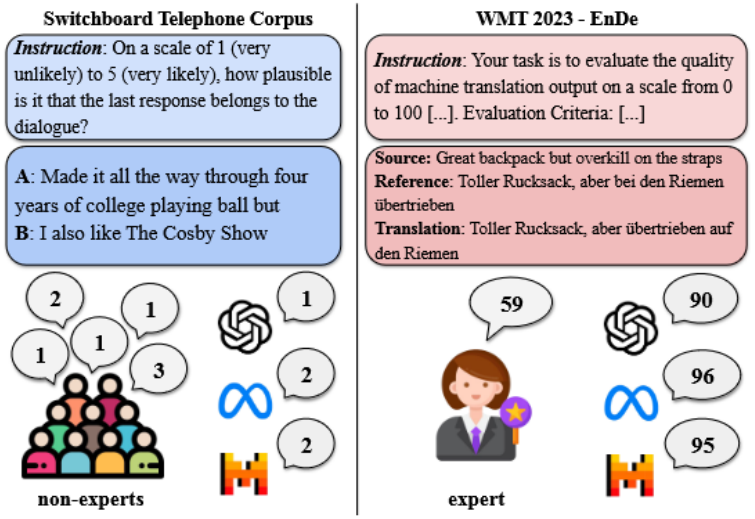
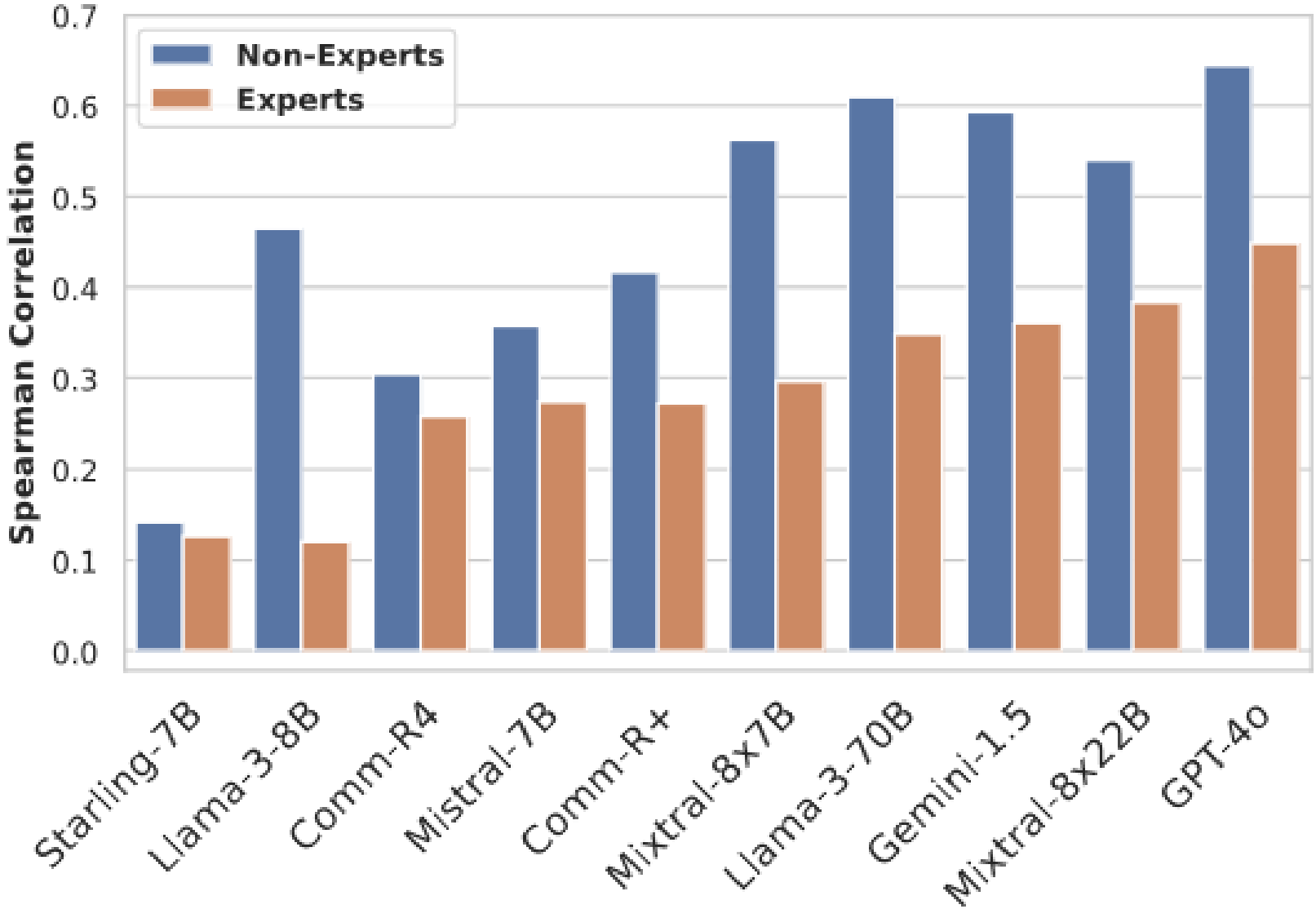


Figure 1: Evaluation by expert and non-expert human annotators and by LLMs for two tasks involving human-generated (left) and machine-generated text (right).


Non-Experts
Experts
Spearman Correlation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Starling-7B
Llama-3-8B
Comm-R4
Mistral-7B
Comm-R+
Mixtral-8x7B
Llama-3-70B
Gemini-1.5
Mixtral-8x22B
GPT-4o

Pragmatic view

- Debate perhaps misguided
  - Humans should remain gold evaluation method
  - But development needs automatic metrics
  - More metrics generally good

→ LLMs instead/in addition to BLEU? Yes please!

→ LLMs as feedback during development? Yes!

→ LLMs as only final evaluation metric? No

Outline

1. Evaluation generics
    1. Hardness
    2. Stages and metrics
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. **Case study: Evaluating factual knowledge**
5. Discussion
    - Group work

Case study: Factual knowledge

- How much factual knowledge do LLMs hold? How accurate is it?
  - Factual knowledge:
    - specific, objective pieces of knowledge about the world
    - *Awards, capitals, Premier league players, …*
    - The kind of knowledge typically found in KBs like Wikidata, Yago, DBpedia
- Question raised at least since 2019
- No conclusive answer or benchmarking technique to date

Evaluation 1: LAMA probe

- Sampled 100k fact triples from Wikidata and Google-RE corpora
  - *(Paris, capitalOf, France)*
  - *(Mammootty, native language, Malayalam)*
- Prompted LLM with verbalized subject-predicate-pair
  - *Paris is the capital of [MASK]*
  - *The native language of Mammootty is [MASK]*
- Evaluated whether sampled object is predicted first (P@1)
- Result: 10-32% P@1 for BERT-large

[Language Models as Knowledge Bases? Petroni et al., 2019]

Problem 1:

- What if more than one object is correct?
- Sampled:
  - *(Germany, sharesBorder, France)*
- LLM-predicted
  - *(Germany, sharesBorder, Poland)*
- P@1 = 0!

Problem 2:
Most subjects have no object

- By sampling triples from existing resources, we only test on what is already known.
- Never test on unknowns/inapplicable relations

- *(Switzerland, officialCapital, -)*
- *(Einstein, doctoralAdvisorOf, -)*
- *(TU Dresden, stocksTradedAt, -)*

“Availability bias” (Tversky and Kahnemann, 1970)

Problem 3: Predictive value for downstream use case

- Evaluation on existing knowledge tells us about reconstruction potential of existing knowledge
  - *hasCapital – 95% recall at precision=90%*
  - *profession – 2% recall at precision=90%*

- But there are 200 countries in the world, which are exhaustively covered online
- There are billions of people….

Problem 4: Things, not strings

- The leader of Jamestown was *John Smith*

- John Smith (anatomist and chemist) (1721–1797), professor of anatomy and chemistry at the University of Oxford, 1767–1797
- John Blair Smith (1764–1799), president of Union College, New York
- John Smith (Cambridge, 1766), vice chancellor of the University of Cambridge, 1766 until 1767
- John Smith (astronomer) (1711–1795), Lowndean Professor of Astronomy and Master of Caius
- John Smith (antiquary) (1747–1807), Scottish antiquary and Gaelic scholar
- John Smith (professor of languages) (1752–1809), professor of languages at Dartmouth College
- John Augustine Smith (1782–1865), president of the College of William and Mary, 1814–1826
- John Smith (botanist) (1798–1888), curator of Kew Gardens
- John Smith (physician) (c.1800–1879), Scottish physician specialising in treating the insane
- John Alexander Smith (physician) (1818–1883), Scottish physician, antiquarian, archaeologist and ornithologist
- J. Lawrence Smith (chemist) (1818–1883), American doctor and chemist
- John Smith (dentist) (1825–1910), founder of Edinburgh's School of Dentistry
- John Campbell Smith (1828–1914), Scottish writer, advocate and Sheriff-Substitute of Forfarshire
- John Donnell Smith (1829–1928), biologist and taxonomist
- John McGarvie Smith (1844–1918), Australian metallurgist and bacteriologist
- John Bernhardt Smith (1858–1912), American entomologist
- John Alexander Smith (1863–1939), British Idealist philosopher
- John Maynard Smith (1920–2004), geneticist
- John Cyril Smith (1922–2003), leading authority on English criminal law
- John Derek Smith (1924–2003), Cambridge molecular biologist
- John Smith (sociologist) (1927–2002), English sociologist

(100 more)

Problem 5: How good is the ground truth?

- Participants in almost all LM-KBC benchmarks:
  - *“In some cases, the ground truth was incomplete/outdated/incorrect“*
- KGs are inherently incomplete
  - *E.g., Wikidata contains only 40/70 members of Dresden‘s city council*

https://reasonator.toolforge.org/?q=Q5123134&lang=en

- Not in ground truth does not mean false!

Evaluation 2: Extrapolation

- Evaluate performance on **random** triples in Wikidata
- Extrapolate to subjects that do not have values in specific relations, but should

Relation	$cardinality^{WD}$	#missing facts	high accuracy predictions(%)	accuracy	#addable facts	growth factor
nativeLanguage	264,778	7,871,085	86%	82%	5,550,689	20.96
spokenLanguage	2,148,775	7,090,119	77%	82%	4,476,701	2.08
headquarteredIn	409,309	55,186	8%	82%	3,443	0.008
developedBy	42,379	29,349	2%	62%	363	0.01
producedBy	123,036	31,239	0.8%	22%	55	0.0004
LanguageOfFilm	337,682	70,669	37%	76%	19,872	0.06
citizenOf	4,206,684	4,616,601	28%	90%	1,163,383	0.27

Strong assumptions:

- works only for mandatory, single-valued relations
- precision on both sets is similar

[Veseli et al., Evaluating language models for knowledge base completion, 2023]

[Singhania et al., 2023]

Evaluation 3: LM-KBC 2023

- Require participating systems to not just return strings, but Wikidata identifiers
  - Q228024 instead of “John Smith”
- Necessitates the use of external entity disambiguation software
  - As LLMs do not sufficiently memorize IDs

Please create a list of Wikidata statements about Q228024, "John Smith the Explorer". Do not write any text, just Q- and P-codes for subject, predicate, object, like "Q228024 P31 Q5".

```
Q228024 P31 Q5
Q228024 P21 Q6581097
Q228024 P106 Q189290
Q228024 P27 Q145
Q228024 P569 +1580-00-00T00:00:00Z/9
Q228024 P570 +1631-00-00T00:00:00Z/9
Q228024 P119 Q307897
Q228024 P800 Q157206
Q228024 P937 Q1549591
Q228024 P1412 Q1860
Q228024 P39 Q486839
Q228024 P512 Q3918
Q228024 P607 Q881
```

citizenship, UK 😀
gender, male 😀
occupation, military officer 🙂
place of burial, <deleted item> ☹
notable work - Goyencourt (commune in France) ☹
position held, member of parlament ☹
conflict, vietnam ☹

Evaluation 3: LM-KBC 2023 (2)

- 21 relations with 100 train/val/test samples from Wikidata

Relation	Macro Average Precision	Macro Average Recall	Macro Average F1-score
BandHasMember	0.5905	0.6331	0.5838
CityLocatedAtRiver	0.7600	0.6538	0.6792
CompanyHasParentOrganisation	0.6100	0.7650	0.6100
CompoundHasParts	0.9962	1.0000	0.9978
CountryBordersCountry	0.8292	0.7699	0.7937
CountryHasOfficialLanguage	0.9379	0.8821	0.8932
CountryHasStates	0.8048	0.8156	0.8073
FootballerPlaysPosition	0.7100	0.7333	0.7083
PersonCauseOfDeath	0.8000	0.8033	0.7983
PersonHasAutobiography	0.4483	0.4850	0.4583
PersonHasEmployer	0.3533	0.3567	0.3282
PersonHasNoblePrize	1.0000	1.0000	1.0000

Evaluation 4: Web-verifiability instead of KB truth

- Compare with web search results instead of KB
  - Retrieve 5 web search results based on subject and object
- Ask for plausibility in addition to truth
- Use LLM-as-judge

> Can the given RDF be inferred from the given snippet?
> Please choose the correct option based on your answer and return only a or b or c or d:
> a) The RDF statement is true according to the snippet.
> b) The RDF statement is plausible according to the snippet.
> c) The RDF statement is implausible according to the snippet.
> d) The RDF statement is false according to the snippet.

→ GPTKB (105M GPT-4o-mini triples) precision is:
31% True, 61% Plausible, 1% Implausible, 7% False

But: Live web search makes results instable!

[Hu et al., GPTKB: Building Very Large Knowledge Bases from Language Models , 2024]

Factual knowledge evaluation: Open issues

- LLM-generations outpace our ways to evaluate them with KBs and on the web
  - LLMs draw plausible inferences (e.g., name → citizenship)
  - Commercial LLMs trained on undisclosed training corpora
- Precision and recall can be traded
  - Best trade-off depends on use case
- How to do approximate matching
  - John Smith, leader, Jamestown
  - John G. Smith, mayor, Jamestown City

Outline

1. Evaluation generics
    1. Hardness
    2. Stages and metrics
    3. Statistics 101
2. Important benchmarks
    1. MMLU
    2. Math
    3. LMArena
3. LLM-as-judge
4. Case study: Evaluating factual knowledge
5. **Discussion**
    - Group work

Group work

- Discuss the following questions with your neighbor:

1. What makes evaluating LLMs so challenging?
2. What are common ways to evaluate LLMs (stages, metrics)?
3. What are the challenges when trying to evaluate factual knowledge in LLMs?

Take home – Lecture 13

1. LLM evaluation challenging
   - Benchmark saturation, test data leakage, open-ended task formats without appropriate metrics, subtle biases, …
   - Investigating and advancing LLM evaluation is valuable research on its own!

2. Common approach
   - Simplification into multiple-choice format
   - LLM-as-judge for open-ended generation tasks

3. Factual knowledge
   - Static benchmarks struggle to evaluate the actual task
   - Not as simple as reconstructing some existing triples

- Lab today:
  - Project kick-off