A Perspective on Evaluation

Swabha Swayamdipta Assistant Professor, USC Viterbi CS Research Fellow, Simons Institute, Special Program on LMs Apr 4, 2025

The Future of LMs and Transformers:

SCiteroi









Text Summarization

Machine Translation Natural Languag Processii Swabha Swayamdipta

Once upon a time...

Question Answering

> Large Language Models

Image from <u>ted.com</u>

Core





Now..







"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be."

> Problems in language (e.g., Information Extraction) were hard for models in the 90s, so we broke them down into smaller tasks which were easy to evaluate. Now models in NLP are extremely good, yet we still evaluate them on simpler tasks!

> > -Claire Cardie, Cornell University, EMNLP 2020 Keynote



-Lord Kelvin; 1860 approx. Source; Skeptics Stack Exchange





Rethinking Evaluation Benchmarks

Benchmarks vs. real-world tasks



50





A Survey on Evaluation of Large Language Models

YUPENG CHANG* and XU WANG*, School of Artificial Intelligence, Jilin University, China JINDONG WANG[†], Microsoft Research Asia, China YUAN WU[†], School of Artificial Intelligence, Jilin University, China LINYI YANG, Westlake University, China KAIJIE ZHU, Institute of Automation, Chinese Academy of Sciences, China HAO CHEN, Carnegie Mellon University, USA XIAOYUAN YI, Microsoft Research Asia, China CUNXIANG WANG, Westlake University, China YIDONG WANG, Peking University, China WEI YE, Peking University, China YUE ZHANG, Westlake University, China YI CHANG, School of Artificial Intelligence, Jilin University, China PHILIP S. YU, University of Illinois at Chicago, USA QIANG YANG, Hong Kong University of Science and Technology, China XING XIE, Microsoft Research Asia, China

> What to evaluate (Sec. 3)

LLM Evaluation mostly focuses on creating new benchmarks











8

Swabha Swayamdipta

Figure 3: HLE consists of 2,700 exam questions in over a hundred subjects, grouped into high level categories here. We provide a more detailed list of subjects in Appendix B.3.



Current LLM evaluation is done on a few popular benchmarks

- Mostly coding and math
 - Perhaps important building blocks for a lot of analytical tasks?
- Most work on benchmarking release only a test set (available to all)
- Many come with no validation set, or a very few validation examples
- LLM developers are incentivized to hill climb on popular benchmarks...

General	MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024b), IFEval (Zhou et al., 2023)
Math and reasoning	GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), GPQA (Rein et al., 2023), ARC-Challenge (Clark et al., 2018)
Code	HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), HumanEval+ (Liu et al., 2024a), MBPP EvalPlus (base) (Liu et al., 2024a), MultiPL-E (Cassano et al., 2023)
Multilinguality	MGSM (Shi et al., 2022), Multilingual MMLU (internal benchmark)
Tool-use	Nexus (Srinivasan et al., 2023), API-Bank (Li et al., 2023b), API-Bench (Patil et al., 2023), BFCL (Yan et al., 2024)
Long context	ZeroSCROLLS (Shaham et al., 2023), Needle-in-a-Haystack (Kamradt, 2023) InfiniteBench (Zhang et al., 2024)

Table 16 Post-training benchmarks by category. Overview of all benchmarks we use to evaluate post-trained Llama 3 models, ordered by capability.

	Model	Prompt	MMLU	GPQA	MATH	HumanEval	MGSM ^[1]	DROP ^[1] (F1, 3- shot)
--	-------	--------	------	------	------	-----------	---------------------	---

3.1. DeepSeek-R1 Evaluation

	Benchmark (Metric)	Claude-3.5- Sonnet-1022	GPT-40 0513	DeepSeek V3	OpenAI 01-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
	MMLU (Pass@1)	88.3	87.2	88.5	85.2	91.8	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	92.9
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	84.0
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	92.2
English	IF-Eval (Prompt Strict)	86.5	84.3	86.1	84.8	-	83.3
English	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	75.7	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	47.0	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	82.5
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	87.6
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	92.3
	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	65.9
Code	Codeforces (Percentile)	20.3	23.6	58.7	93.4	96.6	96.3
	Codeforces (Rating)	717	759	1134	1820	2061	2029
	SWE Verified (Resolved)	50.8	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	61.7	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	79.8
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	97.3
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	78.8
	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	92.8
Chinese	C-Eval (EM)	76.7	76.0	86.5	68.9	-	91.8
	C-SimpleQA (Correct)	55.4	58.7	68.0	40.3	-	63.7



Benchmarks vs. real tasks

- We have evidence language models are getting better and better at math, inference, coding and many logical / analytical tasks
- How about we go bold and test them directly on real tasks with real stakes?
- Caveat: this will involve humans, and much manual labor



Swabha Swayamdipta



-Claire Cardie, Cornell University

Some of the content in the following slides contains sensitive content and can be upsetting







Need AI help for local politicians to understand public attitudes on homelessness to frame the most resonant message to inform public policy.



School for Communication and Journalism

Easy peasy, we've got these amazing language models - let me do this for you in a couple of weeks, tops



@mention how about do something for all the homeless veterans and americans on the streets? you'll give míllíons to the ímmígrants crossing our so called secure border but our homeless americans get nothing. You are merely trying to buy votes

government critique



USCViterbi

School of Engineering



~2.5M tweets on homelessness between 2021-2023

Ranjit, ... & Swayamdipta. EMNLP 2024.









Building a test set took more than 7-8 months of laborious manual work



Results in 6.5x speedup in annotation time!

With some effort, language models can be used as assistants for determining variables which could inform communication and public policy

Swabha Swayamdipta

LMs in Communications / Social Work: Informing policies on homelessness



Ranjit, ... & Swayamdipta. EMNLP 2024.











Heard you have been helping out colleagues with language models. I need some help with analyzing suicide reports that help me validate my hypotheses on novel factors for designing suicide interventions.

USCSuzanne Dworak-Peck School of Social Work

How often do victims interact with non-clinical personnel (e.g. legal professionals) in days before their death?

family custody child support battle with her child

No legal interaction

Explicit legal interaction











LMs in Social Work

Manual Codebook Development:

Legal Interactions



Model	\mathcal{G}_0	\mathcal{G}_{expert}	\mathcal{G}_{12}
Meta-Llama-3-70B	0.57	0.78	0.79
Qwen2.5-32B	0.68	0.77	0.77
Qwen2.5-14B	0.63	0.77	0.76

10.4% of 270K suicide narratives had evidence of legal interactions!



Instead of performance alone, speedup in manual labor at the same performance level should be an important metric in real world tasks

Speedup in process: weeks to hours!

Our research was going at the pace of a bullock cart and you came at it with a jet plane!



Ranjit, ... & Swayamdipta. Under Review. 2025.





Public Health: Emergency Response



Real-world tasks are still hard to evaluate directly, need to be broken down. And human involvement is still important

es later	
ead of a	
er.	



Surana, Ye & Swayamdipta. Under Review. 2025.







Evaluator Efforts and Effects on Human Evaluation

- Human evaluators seem to like language model responses
 - especially those with limited training / experience

The Impact of Generative AI on Critical Thinking: Self-Reported **Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers**

Hao-Ping (Hank) Lee Carnegie Mellon University Pittsburgh, Pennsylvania, USA haopingl@cs.cmu.edu

Ian Drosos Microsoft Research Cambridge, United Kingdom t-iandrosos@microsoft.com

Advait Sarkar Microsoft Research Cambridge, United Kingdom advait@microsoft.com

Sean Rintel Microsoft Research Cambridge, United Kingdom serintel@microsoft.com

Nicholas Wilson Microsoft Research Cambridge, United Kingdom niwilson@microsoft.com

Lev Tankelevitch Microsoft Research Cambridge, United Kingdom levt@microsoft.com

Richard Banks Microsoft Research Cambridge Cambridge, United Kingdom rbanks@microsoft.com

> Figure 2: Distribution of perceived effort (%) in cognitive activities (based on Bloom's taxonomy) when using a GenAI tool compared to not using one.

• experts tend to disagree more with LMs, but can often miss important details too











TradNLP: Text Simplification

Source

Complex Sentence

Simplified Sentence

SimpEval2022 On the fifth day of flight, November 20, 2022, at 19:09 UTC, the Orion spacecraft entered the Lunar sphere of influence, thus the Moon's gravitational force became stronger than Earth's relative to the spacecraft.

Dataset	Human Fluency	Human Meaning	Human Simplicity
Newsela	0.153	0.213	0.054
SimpEval	_		
SynthSimpliEval	—	—	_

Swabha Swayamdipta

Prof, you always ask us to look at our data. I found text simplification benchmarks to be very unsatisfactory, can I do something about it?

On the fifth day of flight, November 20, 2022, at 19:09 UTC, the Orion spacecraft entered the Lunar sphere of influence, so the Moon's gravitational force became stronger than Earth's relative to the spacecraft.

Synthetic data and autoevals are great tools for relatively simpler tasks



Human Unified

> 0.228 0.671*

Liu, Nam, Cui & Swayamdipta, Under Review. 2025











Rethinking Benchmark Tasks

- Our notion of task difficulty might have limitations; using real-world problems to guide tasks is as important as collecting "hard" benchmarks
- Quantitative evaluation of language models on real-world tasks is hard, requires humans with domain expertise
 - Task itself needs to be broken down into smaller parts for ease of evaluation
- Over-reliance on a few select benchmarks might give us a false sense of progress
- Hard tasks (no right answers, experts might disagree): Human verification, real data
 - Small tasks (possible right answers): Auto evals, synthetic data





Rethinking Evaluation Frameworks

Test Data Distributions Generation Baselines





Language models are great at generating language



Write With Transformer distil-gpt2 ①

Shuffle initial text

Trigger autocomplete or tab

Why is Skub the greatest product on the market today? I'll give you three great reasons:

1) Skub is the most reliable solution for your personal, professional and personal life. It offers a great combination of quality and availability.

2) Skub is the most versatile solution for all your clients. It also has a great selection of advanced software, features and products.

3) Skub's unique design and its strong functional design allows you to easily customize your products. It provides great products and services with all the latest features. It has been designed for everyone with a clear look, without needing any further extra effort.

But we still evaluate them on classification tasks: fixed (and often unique) answers

ChatGPT ~

Why is skub the greatest product on the market today? I'll give you three great reasons:

- 1. Unparalleled Versatility Whether you need it for home, work, or play, Skub adapts to every situation with ease. No other product offers such a wide range of applications!
- 2. Cutting-Edge Innovation Skub is designed with the latest technology and backed by years of research. It's not just a product—it's a revolution in convenience and efficiency.
- 3. Undeniable Popularity The people have spoken, and Skub reigns supreme. With rave reviews and an ever-growing fan base, it's clear that Skub isn't just a trend—it's the future.

What more could you ask for? Skub is, simply put, the best.

6 7 4 * ? いく ብ







- Ground truth does not exist for long-form generation
 - Humans may disagree
 - Experts may not be available
- LLMs are compared with each other on their generations directly
 - Still mostly focused on coding / reasoning / knowledge-seeking tasks
- Open Platforms for human raters
- Auto Evals

Human evaluation of generated can be hard, but it's still considered the most reliable method today



Swabha Swayamdipta



Figure 17 Human evaluation results for the Llama 3 405B model. Left: Comparison with GPT-4. Middle: Comparison with GPT-40. Right: Comparison with Claude 3.5 Sonnet. All results include 95% confidence intervals and exclude ties.



X Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots

Blog GitHub Paper Dataset Twitter Discord Kaggle Competition

UB) 🔺	Rank (StyleCtrl)	Model 🔺	Arena Score	95% CI 🔺	Votes 🔺	Organization
	3	Gemini-2.0-Flash-Thinking-Exp- 01-21	1384	+5/-5	11949	Google
	2	Gemini-2.0-Pro-Exp-02-05	1379	+4/-7	9872	Google
	1	ChatGPT-40-latest (2025-01-29)	1377	+6/-6	8643	OpenAI
	2	DeepSeek-R1	1361	+9/-9	4190	DeepSeek
	7	Gemini-2.0-Flash-001	1355	+7/-6	7747	Google
	2	01-2024-12-17	1352	+5/-5	13976	OpenAI
	5	<u>o1-preview</u>	1335	+4/-4	33164	OpenAI
	7	Qwen2.5-Max	1332	+6/-6	5968	Alibaba

New Launch! Jailbreak models at RedTeam Arena





Shane, Cameron, or Saxon—who is the ultimate White Lotus alpha?



Swabha Swayamdipta





When considering who might be the "ultimate alpha" among Shane, Cameron, and Saxon from The White Lotus series, I'd need to analyze each character's portrayal:

MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers

Krishna Pillutla¹ Swabha Swayamdipta² Rowan Zellers¹ John Thickstun³ Sean Welleck^{1,2} Yejin Choi^{1,2} Zaid Harchaoui⁴

¹Paul G. Allen School of Computer Science & Engineering, University of Washington ²Allen Institute for Artificial Intelligence ³Department of Computer Science, Stanford University ⁴Department of Statistics, University of Washington

newlywed who becomes obsessed with getting the Armond in a confrontation. His privilege and e of aggressive dominance.

althy, manipulative finance bro who exerts control even seducing his friend's wife. His confidence and lack social dominance.

in Season 3 set in Thailand. As my knowledge cutoff is

October 2024, I don't have detailed information about his character traits or behaviors in the show.

Based on the first two seasons, both Shane and Cameron exhibit toxic "alpha" behaviors in

Subjectivity in evaluation of long-form text is a feature, not a bug







E CM DOITICS SCOTUS Congress Facts First 2020 Elections

Iran: No signing final nuclear deal unless economic sanctions are lifted

By Faith Karimi and Deirdre Walsh, CNN Indated 3:50 PM EDT. Thu April 9, 2015





(CNN) — Iran will sign a final nuclear agreement only if economic sanctions against the nation are removed on the first day of the deal's implementation, President Hassan Rouhani said Thursday.

Ayatollah Ali Khamenei, the Islamic republic's supreme leader, meanwhile, told state-run media outlets he is neither in favor nor against the proposed deal because it isn't final, and he's not certain it will become binding because he has "never been optimistic about negotiations with the U.S."

Six world powers and Iran reached a preliminary deal last week that aims to limit Tehran's nuclear program in exchange for lifting economic sanctions.

The United States, however, has stressed that if a final deal is reached with Iran, the removal of any sanctions will come in phases.

But work on the agreement isn't finished.

Negotiators from Iran and the United States, China, Germany, France, Britain and Russia have until June 30 to come up with a final deal.

Supreme leader speaks

Khamenei said he supports the negotiators, but in several not-so-subtle shots at the United States, noted it is too soon to celebrate the proposed deal.

"Everything lies in the details," Khamenei said in a Thursday address, according to Press

Summary

Model B Summary





Models A and B focus on different content, but is one necessarily *better*? It's subjective!



Model A Summary

The Russian president Vladimir Putin held a concert in the ruins of Palmyra to show that Russia is a force for good, however many Western officials remain suspicious of Russia's intentions.

The above article discusses a concert that was conducted by Valery Gergiev in Palmyra's Roman theatre in order to protest the barbarism and violence exhibited by Islamic State militants.

The article discusses a concert that was held in Palmyra by the Mariinsky Symphony Orchestra, which was led by Valery Gergiev.

The article discusses a concert that was put on by the Mariinsky Symphony Orchestra in Palmyra's Roman Theatre in order to protest the barbarism and violence exhibited by Islamic State militants who had used the city's Roman amphitheatre to execute prisoners.

The above article discusses a concert that was held in Palmyra by the Mariinsky Symphony Orchestra, led by Valery Gergiev, in order to protest against the barbarism and violence exhibited by Islamic State militants.

Intrinsic Similarity or Self-Alignment



Model B Summary

A Russian conductor has staged a classical concert in the ruins of Palmyra, one of the cities recaptured from Islamic State (IS) militants in Syria.

A Russian conductor has staged a classical concert in the ruins of the ancient Syrian city of Palmyra, which was recently recaptured from Islamic State militants.

A Russian conductor has led a classical concert in the ruins of Palmyra, one of the ancient cities seized by Islamic State (IS) militants in Syria.

A Russian conductor has staged a classical concert in the ruins of Palmyra, one of the ancient cities seized by Islamic State (IS) militants in Syria.

A Russian conductor has staged a classical concert in the ruins of Palmyra, a day after Syrian government forces recaptured the city from Islamic State militants.

Extrinsic Similarity or Cross-Alignment





Alignment

 $\mathscr{A}^{l}_{A,B}$ -

 $A = B \rightarrow Self-Alignment$

Separability

Intuition: How much cross-variability (cross alignment) is there between the output generations of two models being compared, adjusting for the baseline level of variability we see within each model's own output generations (self-alignment)?

Swabha Swayamdipta

Similarity metric, e.g. BERTScore

$$= \frac{1}{K^2} \sum_{j=1}^{K} \sum_{l=1}^{K} s\left(\hat{\mathbf{y}}_A^{i,j}, \hat{\mathbf{y}}_B^{i,l}\right)$$

$$\delta_{A,B}^{i} = \max\left(\mathscr{A}_{A,A}^{i}, \mathscr{A}_{B,B}^{i}\right) - \mathscr{A}_{A,B}^{i}$$

Ghosh, Srinivasan & Swayamdipta, EMNLP Find. 2024







Low SEPARABILITY Test Instance

Ethan: who's going to see solstafir? **Noah:** solstafir! when? **Noah:** i had no idea they were playing Archie: I'm not :D. **Leo:** i'm going ofc! Ethan: 21.11 **Noah:** that's a wednesday... ehh...

• • •









More dissimilar models have higher mean separability







Verifying Separability

- Rating consistency: How faithfully do raters choose the same model across generations for the same input?
- Higher separability should correspond to higher consistency

Are the outputs corresponding to these inputs harder to distinguish by humans?





Separability is correlated with rating consistency





Separability Range



Human Evaluators



(0.0, 0.2] (0.2, 0.4] (0.4, 0.6] (0.6, 0.8] Separability Range













Applying Separability to ELO ratings

- ELO is a popular method to rank models, e.g. chatbotarena
- $ELO'_{A} = ELO_{A} + K^{i}(S^{i}_{A} E^{i}_{A}),$
 - ELO_A is the original rating
 - S_A^i is the outcome of the comparison with instance i,
 - E_A^i is the expected win probability • K^i is a weighting factor

α We consider $K_{SEP}^i = K^i$ $-\beta(\delta^i_{AB})$ $1 + \exp \left(-\frac{1}{2} + \frac{1}{2} + \frac{1}$

• T, β, α are all hyperparameters

Swabha Swayamdipta



This ensures that rankings are better calibrated, and we do not overly rely on models which might not be as effective

Ghosh, Srinivasan & Swayamdipta, EMNLP Find. 2024









Limitations of Separability

- Dependent on model pair
- Requires inference time scaling, but for evaluation...



Does not reveal anything about inputs (prompts), to the best of our knowledge



Ghosh, Srinivasan & Swayamdipta, EMNLP Find. 2024





From self-alignment to consensus

- If samples have a high degree of variance, and individual samples are all valuable, could we get more valuable generations by simply "merging" them?
 - Outliers may contain hallucinations
- Different from inference-time scaling approaches which pick a single generation (e.g. self consistency, or LLM monkeys) or produce a very long generation (reasoning chains)

however many Western officials remain suspicious of Russia's intentions.

protest the barbarism and violence exhibited by Islamic State militants.

Valery Gergiev.

amphitheatre to execute prisoners.

Gergiev, in order to protest against the barbarism and violence exhibited by Islamic State militants.

- The Russian president Vladimir Putin held a concert in the ruins of Palmyra to show that Russia is a force for good,
- The above article discusses a concert that was conducted by Valery Gergiev in Palmyra's Roman theatre in order to
- The article discusses a concert that was held in Palmyra by the Mariinsky Symphony Orchestra, which was led by
- The article discusses a concert that was put on by the Mariinsky Symphony Orchestra in Palmyra's Roman Theatre in order to protest the barbarism and violence exhibited by Islamic State militants who had used the city's Roman
- The above article discusses a concert that was held in Palmyra by the Mariinsky Symphony Orchestra, led by Valery

Ghosh, Yauney,, Warraich & Swayamdipta, In Prep.





Inference-time Scaling to obtain Consensus Baselines



Swabha Swayamdipta

Ghosh, Yauney,, Warraich & Swayamdipta, In Prep.







Consensus: Preliminary Results

- Can serve as a baseline!
- Remember, we may not have a verifier in long-form generation
- Based on the assumption that repeatedly generated factoids tend to be those that the model is confident about
- Performs better than approaches which sample a single generation

Swabha Swayamdipta



Factscore

Average FactScore of 5

GPT-40 Consensus

Minimum Bayes Risk Decoding with BERTScore

Consensus (>=0.1) agreement

Consensus (>=0.3) agreement

Consensus (>=0.5) agreement

Consensus (>=0.7) agreement

Consensus (>=0.9) agreement

Ghosh, Yauney,, Warraich & Swayamdipta, In Prep.



Rethinking Test Data Distributions

- Evaluating long form generation is much harder than evaluating classification
- The search for an ideal test data distribution is a wild goose chase
- However, we need to contextualize benchmark performance given that our test sets are far from perfect
- We need more intrinsic baselines in long-form generation (like the random baseline in classification)

Swabha Swayamdipta

Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics

Swabha Swayamdipta[†] Roy Schwartz^{‡*} Nicholas Lourie[†] Yizhong Wang^{\diamond} Hannaneh Hajishirzi^{$\dagger \diamond$} Noah A. Smith^{$\dagger \diamond$} Yejin Choi^{$\dagger \diamond$}

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA [‡]The Hebrew University of Jerusalem, Israel Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

> **Compare without Despair: Reliable Preference Evaluation with Generation SEPARABILITY**

> Sayan Ghosh Tejas Srinivasan Swabha Swayamdipta Thomas Lord Department of Computer Science, University of Southern California {ghoshsay, tejas.srinivasan, swabhas}@usc.edu

Annotation Artifacts in Natural Language Inference Data

Suchin Gururangan \star^{\diamond} Swabha Swayamdipta \star^{\heartsuit} Omer Levy* Roy Schwartz** Samuel R. Bowman[†] Noah A. Smith*

[◊] Department of Linguistics, University of Washington, Seattle, WA, USA [♡] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA * Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA Allen Institute for Artificial Intelligence, Seattle, WA, USA [†] Center for Data Science and Department of Linguistics, New York University, New York, NY, USA

Understanding Dataset Difficulty with V-Usable Information

Kawin Ethayarajh¹ Yejin Choi²³ Swabha Swayamdipta²

Adversarial Filters of Dataset Biases

Ronan Le Bras¹ **Swabha Swayamdipta**¹ **Chandra Bhagavatula**¹ **Rowan Zellers**¹² **Matthew E. Peters**¹ Ashish Sabharwal¹ Yejin Choi¹²



Evaluating Language Models is Hard, Hard, Hard



Rethinking the Target of Evaluation

Assessing LLMs through Numerical Outputs, Activations and Parameters

-0.32643065, -0.12308089, -0.2873811 , -0.99628943, -0.2503798 , 0.24311952, 0.5662387 , 0.17282294, -0.1109335 , 0.15209009, 0 47017908, -0.19270805,



Language model outputs go beyond just tokens



LM output probabilities (or logits) hold a lot of information

Finlayson, Ren & Swayamdipta, COLM 2024



Language Models have a Softmax Bottleneck

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of \mathbb{R}^{ν} and Δ_{ν} , respectively
- This final layer is thus low-rank, since $v \gg d$ Yang et al., ICLR 2018; Finlayson et al., ICLR 2024
- A collection of *d* linearly independent outputs $\mathbf{p}^1, \mathbf{p}^2, \dots, \mathbf{p}^d \in \Delta_v$ from the model will form a basis for the model's image
- We call the image of the model, i.e. LM outputs in either **l** or **p**, the model signature



All LM outputs can be expressed as a unique linear combination of these *d* outputs

Finlayson, Ren & Swayamdipta, COLM 2024





Model signatures reveal LM sizes and identify outputs

- Even different checkpoints from the same LM have largely disjoint model signatures
 - access to a set of LMs and without knowing the exact inputs to the model.



GPT-3.5-Turbo has hidden dimension close to 4096 and is likely a 7B model! Note: in Feb 2024

• Possible to determine precisely which LM produced a particular output, using only API



Potential Implications of Model Internals

- Can language models be evaluated without language?
- Model Signatures
 - How do these change during training, from checkpoint to checkpoint? • Can they reveal newer axes of comparison, such as vulnerability to attacks?
- Could other model internals hold clues for model capabilities? • e.g. Unembedding layers, Representations due to Layer Norms

 - Latent representations, such as "skills"

A Theory for Emergence of Complex Skills in Language Models

Sanjeev Arora Princeton University*

Anirudh Goyal Google DeepMind





Nathan Godey^{1,2}, Éric de la Clergerie¹ & Benoît Sagot¹ ¹ Inria Paris, ² Sorbonne Université Paris, France nathan.godey@inria.fr



The Future of LMs: A Perspective on Evaluation

- Evaluation will ultimately determine the future of LMs
- We need to question age-old assumptions about evaluations
 - Check out Ben Recht's blog: <u>https://</u> <u>www.argmin.net/</u>
- Current trends which may lead to diminishing returns
 - Fixation on a handful of popular benchmarks
 - Vibes-only evaluation







Work by Jaspreet Ranjit, Matt Finlayson, Sayan Ghosh, Joseph Liu, Risha Surana, Greg Yauney, Xinyue Cui, Brihi Joshi, Atharva Kulkarni, Qinyuan Ye, Yoonsoo Nam, Justin Cho and others

Swabha Swayamdipta

USCViterbi Natural Language Processing



swabhas@usc.edu

