Evaluation and Benchmarking

Part 1: Background and standard NLP benchmarks

Vladislav Mikhailov

Language Technology Group, University of Oslo



Table of contents



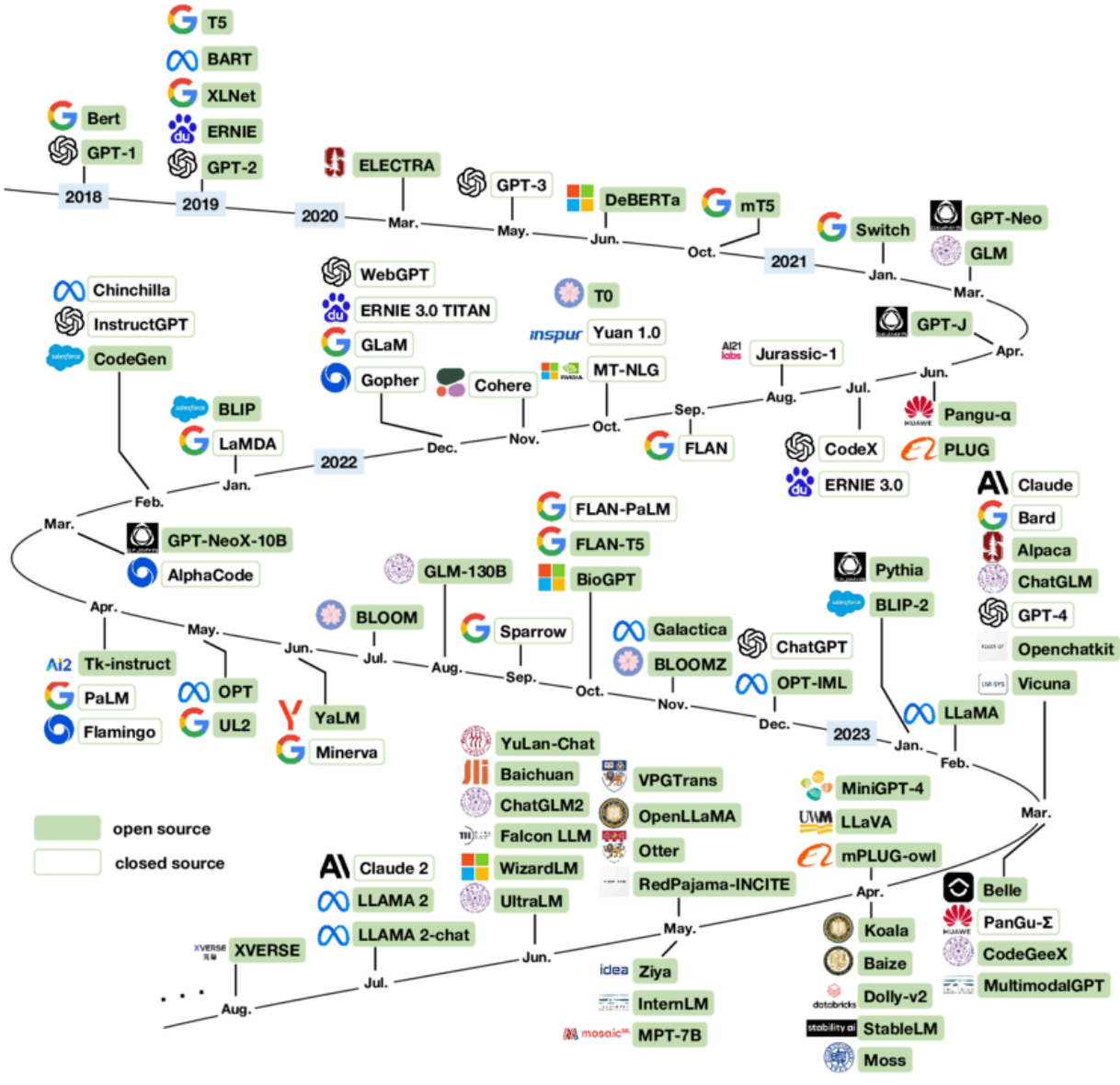
1. Background

2. Standard NLP Benchmarks

3. References

Background

A brief overview of language models (LMs). Source: Gao et al. (2023).



Background

How can we objectively evaluate deep learning models?

- Benchmarking as a standard approach to comparing systems w.r.t. various criteria, e.g.:
  - performance;
  - computational efficiency;
  - safety.
- Benchmarks help track progress in different applied areas of deep learning and accelerate model development.

What is a benchmark?

- A *benchmark* is a collection of datasets, performance metrics, and an aggregation procedure.
- Benchmarks are gamified with public leaderboards, where deep learning models and human annotators are ranked against one another.

Rank	Model	Score	Task_1 (Acc.)	Task_2 (F1)	Task_3 (F1)
1	A	0.95	**0.98**	0.93	**0.95**
2	Human	0.94	**0.98**	**0.96**	0.93
3	B	0.90	0.95	0.95	0.85

Table 1: A toy leaderboard for illustration purposes.

Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].

[1] *paperswithcode.com/sota*



Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.

[1] *paperswithcode.com/sota*



Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.

[1] *paperswithcode.com/sota*

Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.
  - language understanding: how well an LM understands the language.

[1] *paperswithcode.com/sota*



Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.
  - language understanding: how well an LM understands the language.
  - language generation: how well an LM produces the language.

[1] *paperswithcode.com/sota*

Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.
  - language understanding: how well an LM understands the language.
  - language generation: how well an LM produces the language.
  - cross-lingual knowledge transfer: how well an LM transfers the knowledge between languages.

[1] *paperswithcode.com/sota*

Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.
  - language understanding: how well an LM understands the language.
  - language generation: how well an LM produces the language.
  - cross-lingual knowledge transfer: how well an LM transfers the knowledge between languages.
  - fairness: how often an LM reproduces social biases and stereotypes.

[1] *paperswithcode.com/sota*

Benchmarking is a rapidly developing direction.

- There are more than 12,500 benchmarks, with 2,300 of them developed in the NLP field[1].
- Directions on evaluating and benchmarking language models (LMs) in NLP:
  - language modeling: how well an LM estimates a text probability.
  - text representation: how well an LM encodes a text.
  - language understanding: how well an LM understands the language.
  - language generation: how well an LM produces the language.
  - cross-lingual knowledge transfer: how well an LM transfers the knowledge between languages.
  - fairness: how often an LM reproduces social biases and stereotypes.
  - other: probing and application-specific benchmarks.

[1] *paperswithcode.com/sota*

Table of contents



1. Background

2. Standard NLP Benchmarks

3. References

Standard NLP Benchmarks



Language modeling evaluation.

- Word prediction task:
  - WikiText (Merity et al., 2016);
  - LAMBADA (Paperno et al., 2016).
- Performance metric: accuracy.
- E.g.:
  - **Context**: Again, he left that up to you. However, he was adamant in his desire that it remain a private ceremony. He asked me to make sure, for instance, that no information be given to the newspaper regarding his death, not even an obituary.
  - **Target sentence**: I got the sense that he didn't want anyone, aside from the three of us, to know that he'd even ____.
  - **Target word**: died

Standard NLP Benchmarks

Language modeling evaluation.

- Text completion task:
  - StoryCloze (Mostafazadeh et al., 2017);
  - HellaSwag (Zellers et al., 2019).
- Performance metric: accuracy.

Context	Right Ending	Wrong Ending
Sammy's coffee grinder was broken. He needed something to crush up his coffee beans. He put his coffee beans in a plastic bag. He tried crushing them with a hammer.	It worked for Sammy.	Sammy was not that much into coffee.
Gina misplaced her phone at her grandparents. It wasnt anywhere in the living room. She realized she was in the car before. She grabbed her dads keys and ran outside.	She found her phone in the car.	She didnt want her phone anymore.
Sarah had been dreaming of visiting Europe for years. She had finally saved enough for the trip. She landed in Spain and traveled east across the continent. She didn't like how different everything was.	Sarah decided that she preferred her home over Europe.	Sarah then decided to move to Europe.

Figure 1: An example from StoryCloze.

Standard NLP Benchmarks

Language modeling evaluation.

- Estimating a text probability:
  - Penn Treebank (Marcus et al., 1993);
  - One Billion Word (Chelba et al., 2013).
- Performance metric: perplexity.

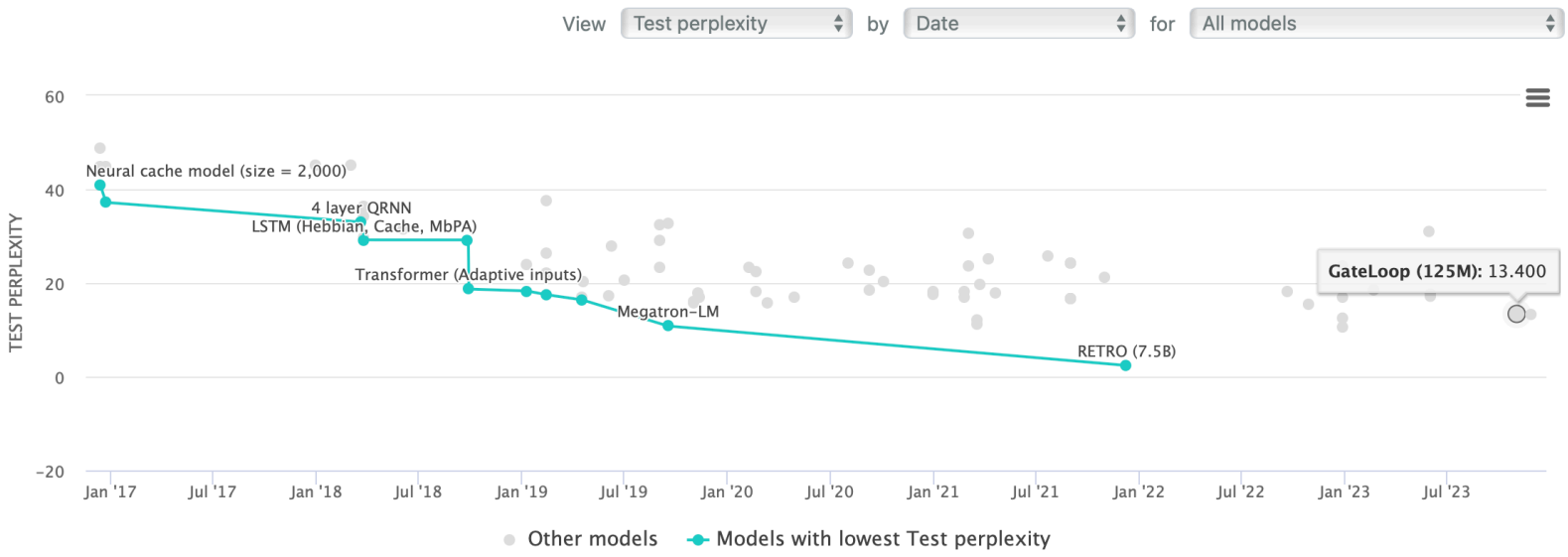


Figure 2: An overview of language model performance scores on WikiText-103.

Standard NLP Benchmarks

Text representation evaluation.

- MTEB (Muennighoff et al., 2023).

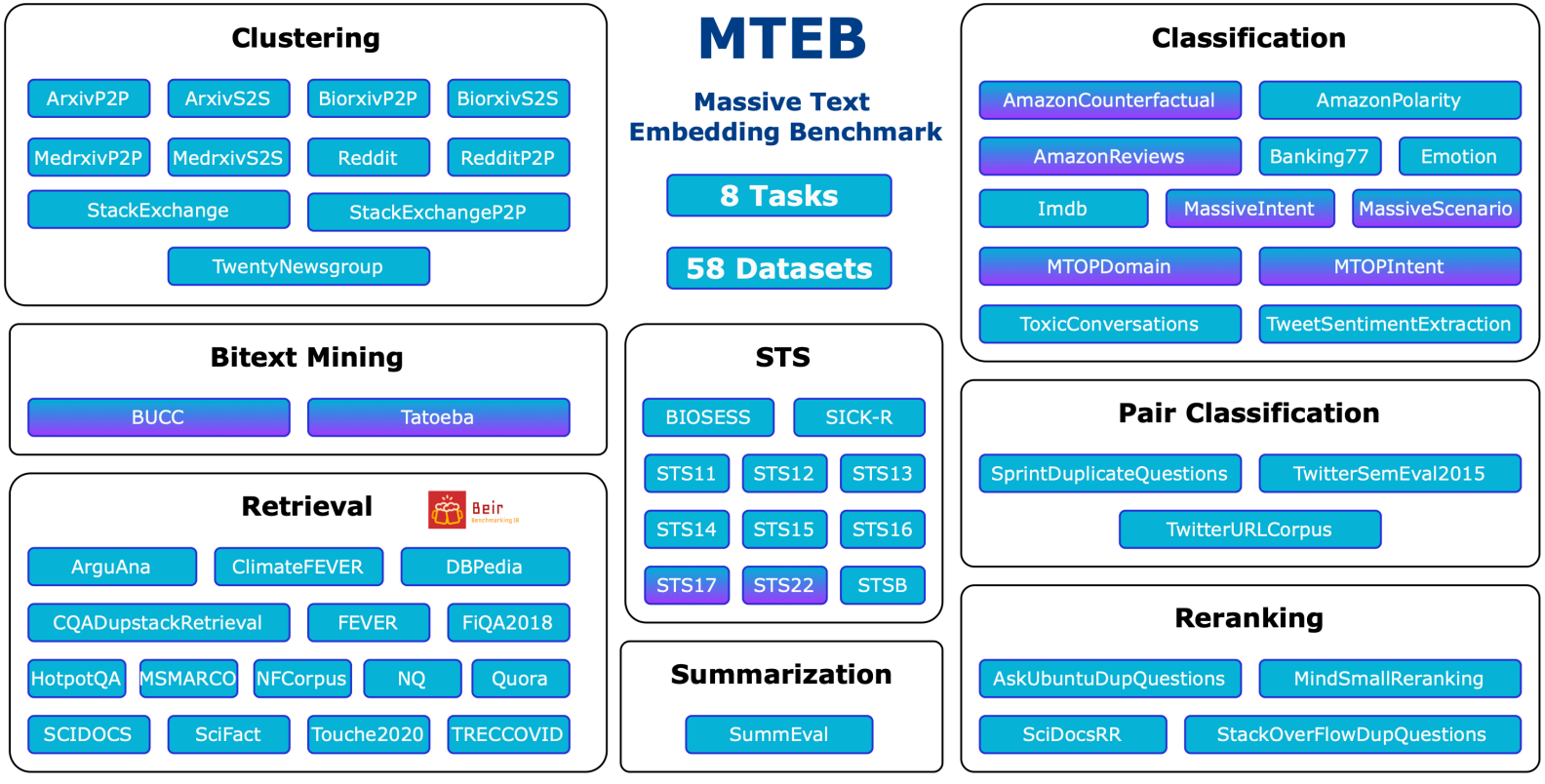


Figure 3: An overview of tasks and datasets in MTEB.

Standard NLP Benchmarks

Natural language understanding evaluation.

- GLUE (Wang et al., 2019b) & SuperGLUE (Wang et al., 2019a):
  - Korean (Park et al., 2021), Polish (Rybak et al., 2020), etc.

ReCoRD	**Paragraph:** *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood* **Query** For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** `US`
RTE	**Text:** *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.* **Hypothesis:** *Christopher Reeve had an accident.* **Entailment:** `False`
WiC	**Context 1:** *Room and board.* **Context 2:** *He nailed boards across the windows.* **Sense match:** `False`
WSC	**Text:** *Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful.* **Coreference:** `False`

Figure 4: A sample of the SuperGLUE tasks.

Standard NLP Benchmarks

Natural language generation evaluation.

- GEM (Gehrmann et al., 2021).

Dataset	Communicative Goal	Language(s)	Size	Input Type
CommonGEN (Lin et al., 2020)	Produce a likely sentence which mentions all of the source concepts.	en	67k	Concept Set
Czech Restaurant (Dušek and Jurčíček, 2019)	Produce a text expressing the given intent and covering the specified attributes.	cs	5k	Meaning Representation
DART (Radev et al., 2020)	Describe cells in a table, covering all information provided in triples.	en	82k	Triple Set
E2E clean (Novikova et al., 2017) (Dušek et al., 2019)	Describe a restaurant, given all and only the attributes specified on the input.	en	42k	Meaning Representation
MLSum (Scialom et al., 2020)	Summarize relevant points within a news article	*de/es	*520k	Articles
Schema-Guided Dialog (Rastogi et al., 2020)	Provide the surface realization for a virtual assistant	en	*165k	Dialog Act
ToTTo (Parikh et al., 2020)	Produce an English sentence that describes the highlighted cells in the context of the given table.	en	136k	Highlighted Table
XSum (Narayan et al., 2018)	Highlight relevant points in a news article	en	*25k	Articles
WebNLG (Gardent et al., 2017)	Produce a text that verbalises the input triples in a grammatical and natural way.	en/ru	50k	RDF triple
WikiAuto + Turk/ASSET (Jiang et al., 2020) (Xu et al., 2016) (Alva-Manchego et al., 2020)	Communicate the same information as the source sentence using simpler words and grammar.	en	594k	Sentence
WikiLingua (Ladhak et al., 2020)	Produce high quality summaries of an instructional article.	*ar/cs/de/en es/fr/hi/id/it ja/ko/nl/pt/ru th/tr/vi/zh	*550k	Article

Figure 5: An overview of the GEM tasks.

Standard NLP Benchmarks

Cross-lingual knowledge transfer evaluation.

- A mixture of language understanding, generation, and sentence retrieval tasks:
  - XGLUE (Liang et al., 2020);
  - XTREME (Hu et al., 2020).
- A multilingual model is finetuned on data in a source language (e.g., English) and evaluated on test data in a target language (e.g., Chinese).

Task	Corpus	|Train|	|Dev|	|Test|	Test sets	|Lang.|	Task	Metric	Domain
Classification	XNLI	392,702	2,490	5,010	translations	15	NLI	Acc.	Misc.
	PAWS-X	49,401	2,000	2,000	translations	7	Paraphrase	Acc.	Wiki / Quora
Struct. pred.	POS	21,253	3,974	47-20,436	ind. annot.	33 (90)	POS	F1	Misc.
	NER	20,000	10,000	1,000-10,000	ind. annot.	40 (176)	NER	F1	Wikipedia
QA	XQuAD	87,599	34,726	1,190	translations	11	Span extraction	F1 / EM	Wikipedia
	MLQA			4,517–11,590	translations	7	Span extraction	F1 / EM	Wikipedia
	TyDiQA-GoldP	3,696	634	323–2,719	ind. annot.	9	Span extraction	F1 / EM	Wikipedia
Retrieval	BUCC	-	-	1,896–14,330	-	5	Sent. retrieval	F1	Wiki / news
	Tatoeba	-	-	1,000	-	33 (122)	Sent. retrieval	Acc.	misc.

Figure 6: An overview of the XTREME tasks.

Standard NLP Benchmarks



Benchmarking is becoming more complex.

- MMLU (Hendrycks et al., 2020): language understanding tasks requiring specialized domain knowledge;

Standard NLP Benchmarks



Benchmarking is becoming more complex.

- MMLU (Hendrycks et al., 2020): language understanding tasks requiring specialized domain knowledge;
- CrossFit (Ye et al., 2021): cross-task few-shot evaluation;

Standard NLP Benchmarks

Benchmarking is becoming more complex.

- MMLU (Hendrycks et al., 2020): language understanding tasks requiring specialized domain knowledge;
- CrossFit (Ye et al., 2021): cross-task few-shot evaluation;
- BigBench (Srivastava et al., 2023): more than 100 NLP tasks;

Standard NLP Benchmarks



Benchmarking is becoming more complex.

- MMLU (Hendrycks et al., 2020): language understanding tasks requiring specialized domain knowledge;
- CrossFit (Ye et al., 2021): cross-task few-shot evaluation;
- BigBench (Srivastava et al., 2023): more than 100 NLP tasks;
- HELM (Liang et al., 2023): user-oriented evaluation scenarios;

Standard NLP Benchmarks



Benchmarking is becoming more complex.

- MMLU (Hendrycks et al., 2020): language understanding tasks requiring specialized domain knowledge;
- CrossFit (Ye et al., 2021): cross-task few-shot evaluation;
- BigBench (Srivastava et al., 2023): more than 100 NLP tasks;
- HELM (Liang et al., 2023): user-oriented evaluation scenarios;
- TURINGBENCH (Uchendu et al., 2021) & M4 (Wang et al., 2023): detecting generated texts in various experimental setups.

Standard NLP Benchmarks

Professional Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk ✗
(B) Branch of the external carotid artery ✗
(C) Branch of the thyrocervical trunk ✓
(D) Tributary of the internal jugular vein ✗

Figure 7: A question from the MMLU's Professional Medicine task.

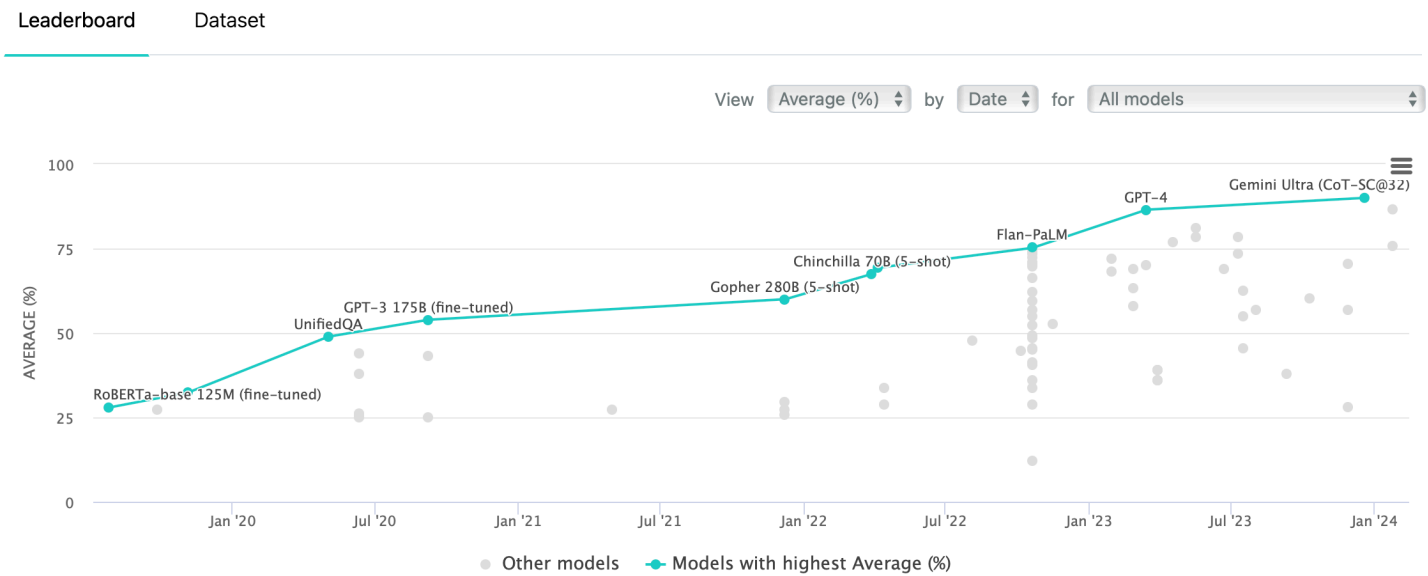
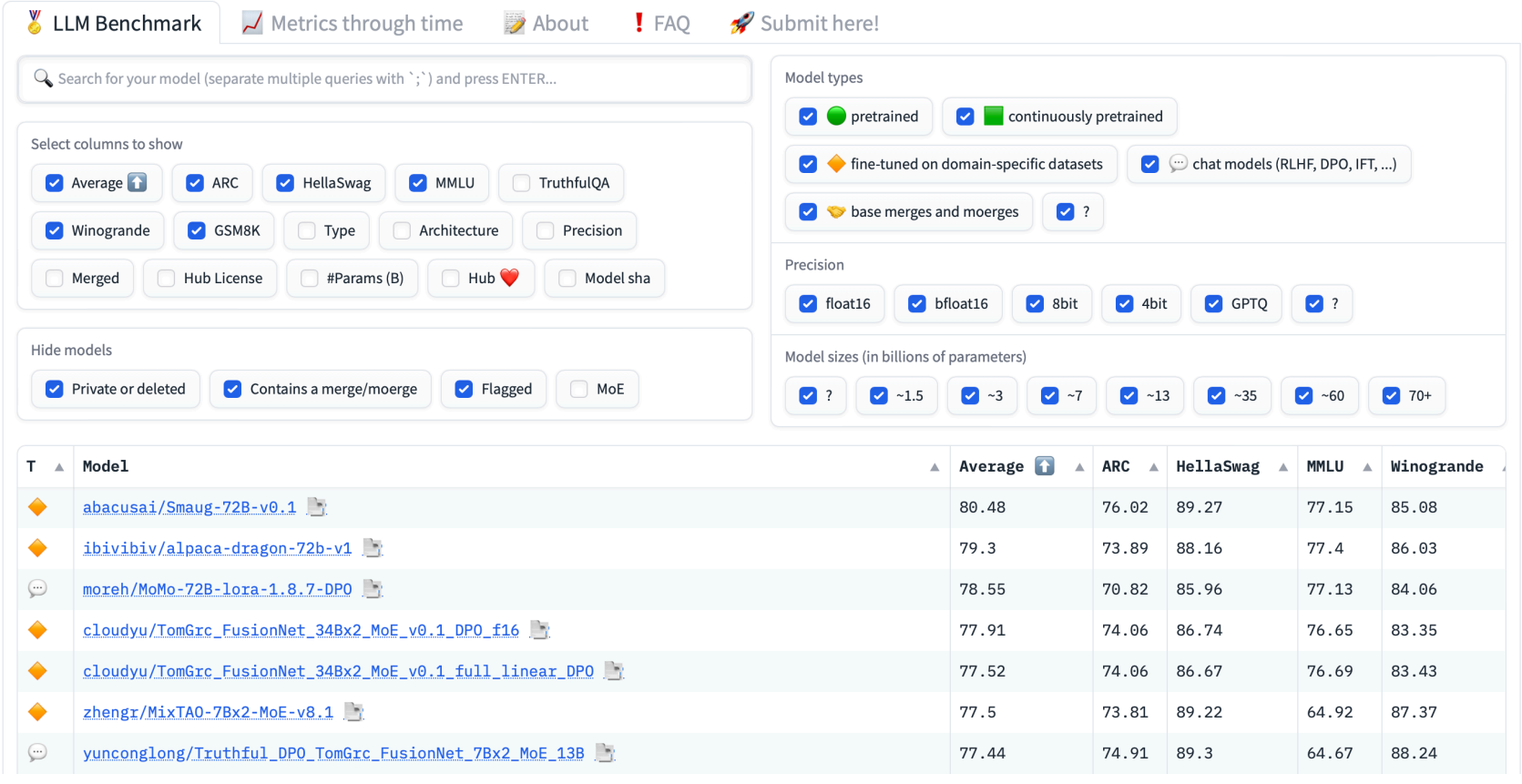


Figure 8: An overview of language model performance scores on MMLU.

Standard NLP Benchmarks



T	Model	Average	ARC	HellaSwag	MMLU	Winogrande
🔶	abacusai/Smaug-72B-v0.1	80.48	76.02	89.27	77.15	85.08
🔶	ibivibiv/alpaca-dragon-72b-v1	79.3	73.89	88.16	77.4	86.03
💬	moreh/MoMo-72B-lora-1.8.7-DPO	78.55	70.82	85.96	77.13	84.06
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_DPO_f16	77.91	74.06	86.74	76.65	83.35
🔶	cloudyu/TomGrc_FusionNet_34Bx2_MoE_v0.1_full_linear_DPO	77.52	74.06	86.67	76.69	83.43
🔶	zhengr/MixTAO-7Bx2-MoE-v8.1	77.5	73.81	89.22	64.92	87.37
💬	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	77.44	74.91	89.3	64.67	88.24

Figure 9: The OpenLLM leaderboard[2].

[2] *hf.co/HuggingFaceH4/open_llm_leaderboard*

Standard NLP Benchmarks

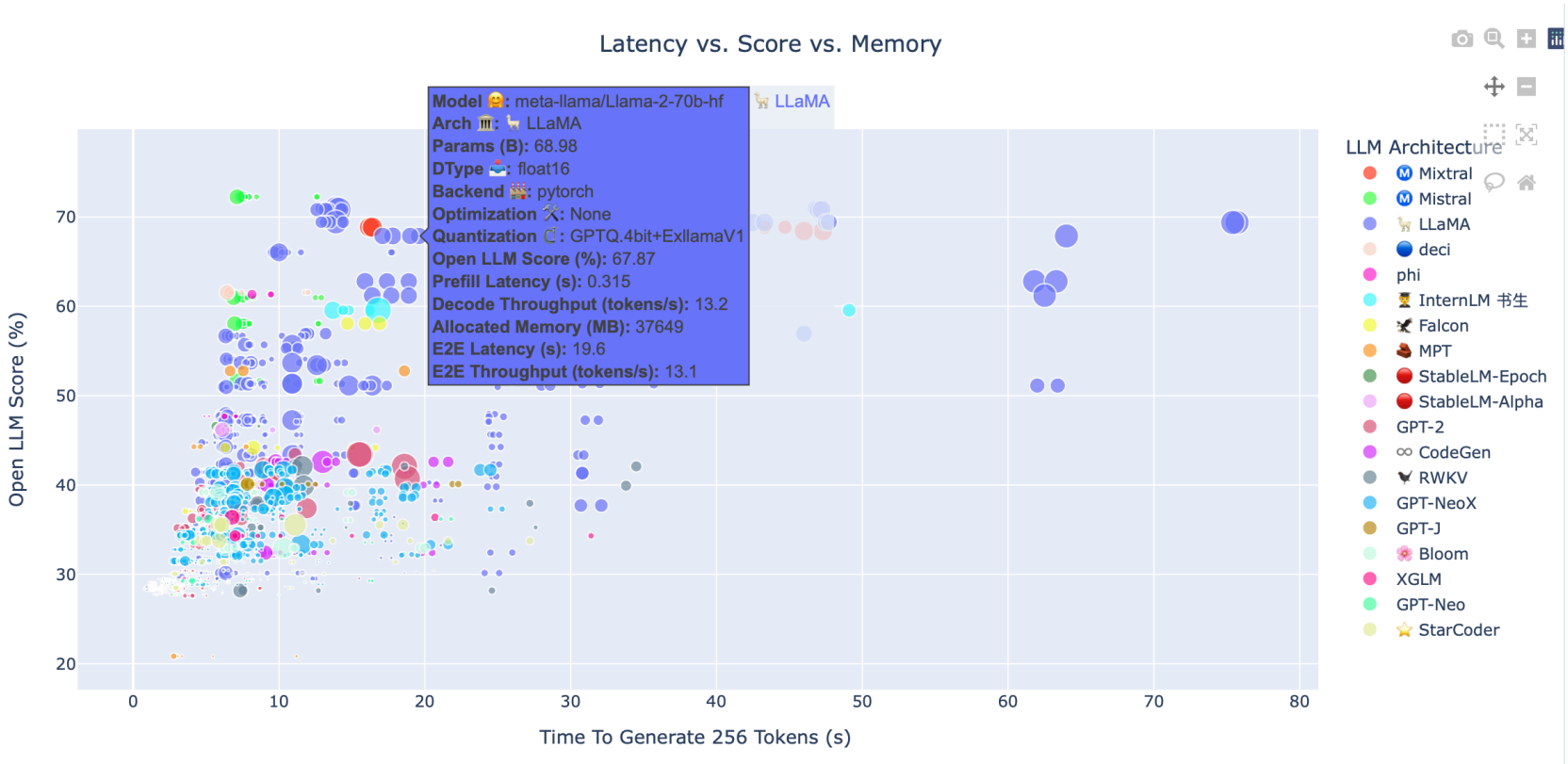


Figure 10: The LLM-perf leaderboard[3].

[3] *hf.co/optimum/llm-perf-leaderboard*

Table of contents



1. Background

2. Standard NLP Benchmarks

3. References

References I

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2013). One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *arXiv preprint arXiv:1312.3005*.

Gao, K., He, S., He, Z., Lin, J., Pei, Q., Shao, J., and Zhang, W. (2023). Examining User-friendly and Open-sourced Large GPT Models: A Survey on Language, Multimodal, and Scientific GPT models. *arXiv preprint arXiv:2308.14149*.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezudo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

References II

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). XTREME: A Benchmark for Evaluating Cross-lingual generalization. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. (2020). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.

References III

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer Sentinel Mixture Models. In *International Conference on Learning Representations*.

Mostafazadeh, N., Roth, M., Louis, A., Chambers, N., and Allen, J. (2017). LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023). MTEB: Massive text embedding benchmark. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, N. Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J. Y., Park, J., Song, C., Kim, J., Song, Y., Oh, T., et al. (2021). Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

References iv

Rybak, P., Mroczkowski, R., Tracz, J., and Gawlik, I. (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Uchendu, A., Ma, Z., Le, T., Zhang, R., and Lee, D. (2021). TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019a). *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., Whitehouse, C., Afzal, O. M., Mahmoud, T., Aji, A. F., and Nakov, P. (2023). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection.

References v

Ye, Q., Lin, B. Y., and Ren, X. (2021). CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.