06.05.24

OPENCAMPUS.sh

Fine-Tuning and Deployment of Large Language Models

MODEL EVALUATION

- **News**
- **Model Evaluation**
- **Project Discussions**
- **Tasks until next week**

NEWS

- Who is doing the news section next week?

W&B CONFIG

```python
import types

# Define hyperparameters
config = types.SimpleNamespace(
    learning_rate=0.001,
    batch_size=32,
    num_epochs=10
    # Add other hyperparameters as needed
)
```

W&B WITH HUGGING FACE

```python
from transformers import Trainer, TrainingArguments
import wandb

# Initialize wandb
wandb.init(project="my-huggingface-project", config=config.__dict__)

# Set up TrainingArguments with wandb integration
training_args = TrainingArguments(
    output_dir="./results",
    learning_rate=config.learning_rate,
    per_device_train_batch_size=config.batch_size,
    num_train_epochs=config.num_epochs,
    report_to="wandb"
    # other arguments...
)
# Create a Trainer with your model, data, and training_args
trainer = Trainer(model=my_model, args=training_args,
                  train_dataset=train_dataset, eval_dataset=eval_dataset)
# Run training
trainer.train()
```

W&B WITH PYTORCH

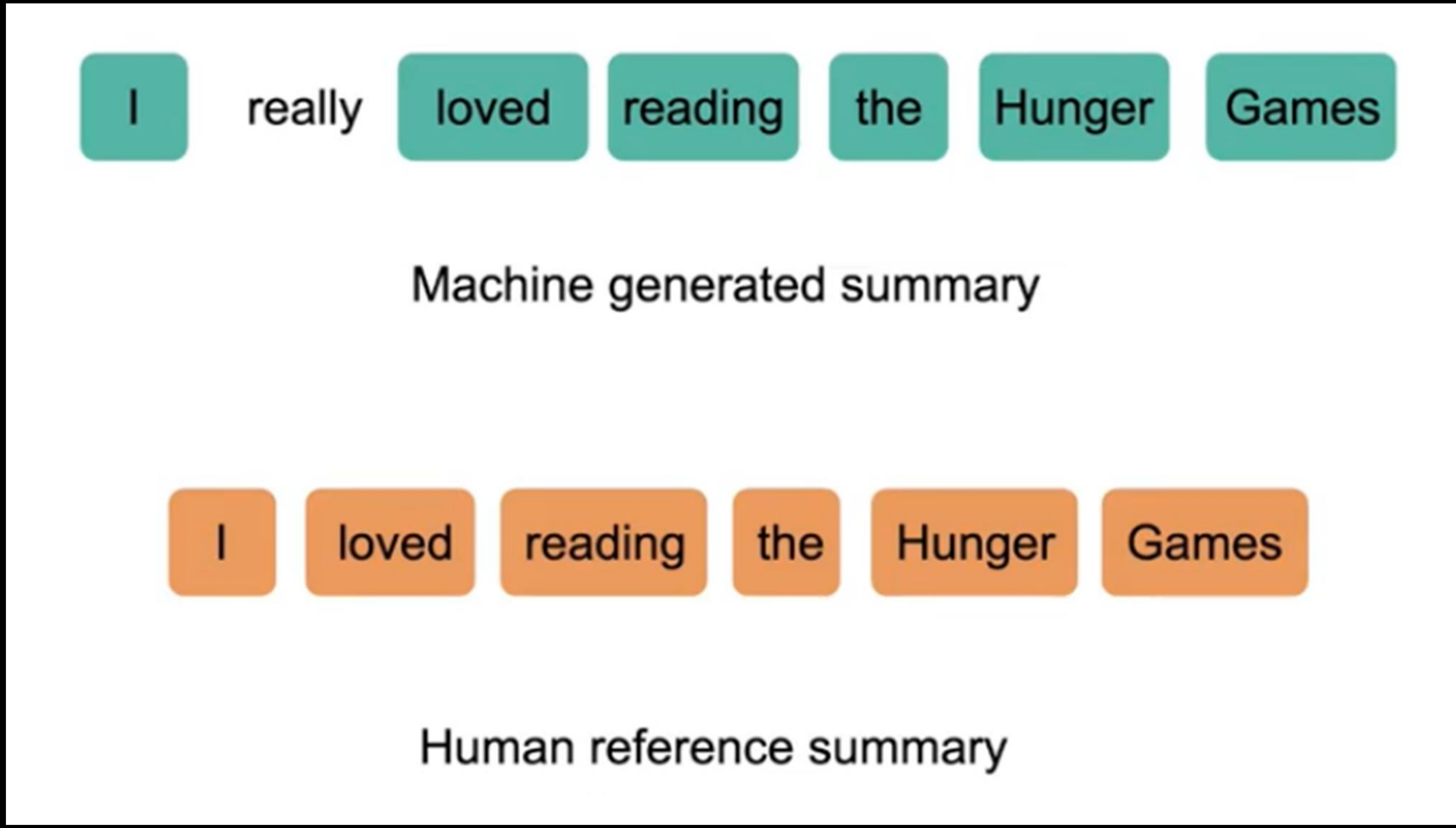
```python
import torch
import wandb

# Initialize wandb
wandb.init(project="my-pytorch-project", config=config.__dict__)

# Define a PyTorch model
model = ...

# Define optimizer with the learning rate from the config
optimizer = torch.optim.Adam(model.parameters(), lr=config.learning_rate)

# Training loop
for epoch in range(config.num_epochs):
    # Training steps
    # ...
    # Log metrics or other information if needed
    wandb.log({"loss": loss})
```

W&B WITH TENSORFLOW

```
import wandb
from wandb.keras import WandbCallback
import tensorflow as tf

# Initialize wandb
wandb.init(project="my-tensorflow-project", config=config.__dict__)

# Create and compile a Keras model
model = ...

# Compile the model with the learning rate from the config
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=config.learning_rate),
              loss='...', metrics=['...'])

# Train the model with WandbCallback
model.fit(x_train, y_train, batch_size=config.batch_size, epochs=config.num_epochs,
callbacks=[WandbCallback()])
```

MODEL EVALUATION

ROUGE- 1 SCORE



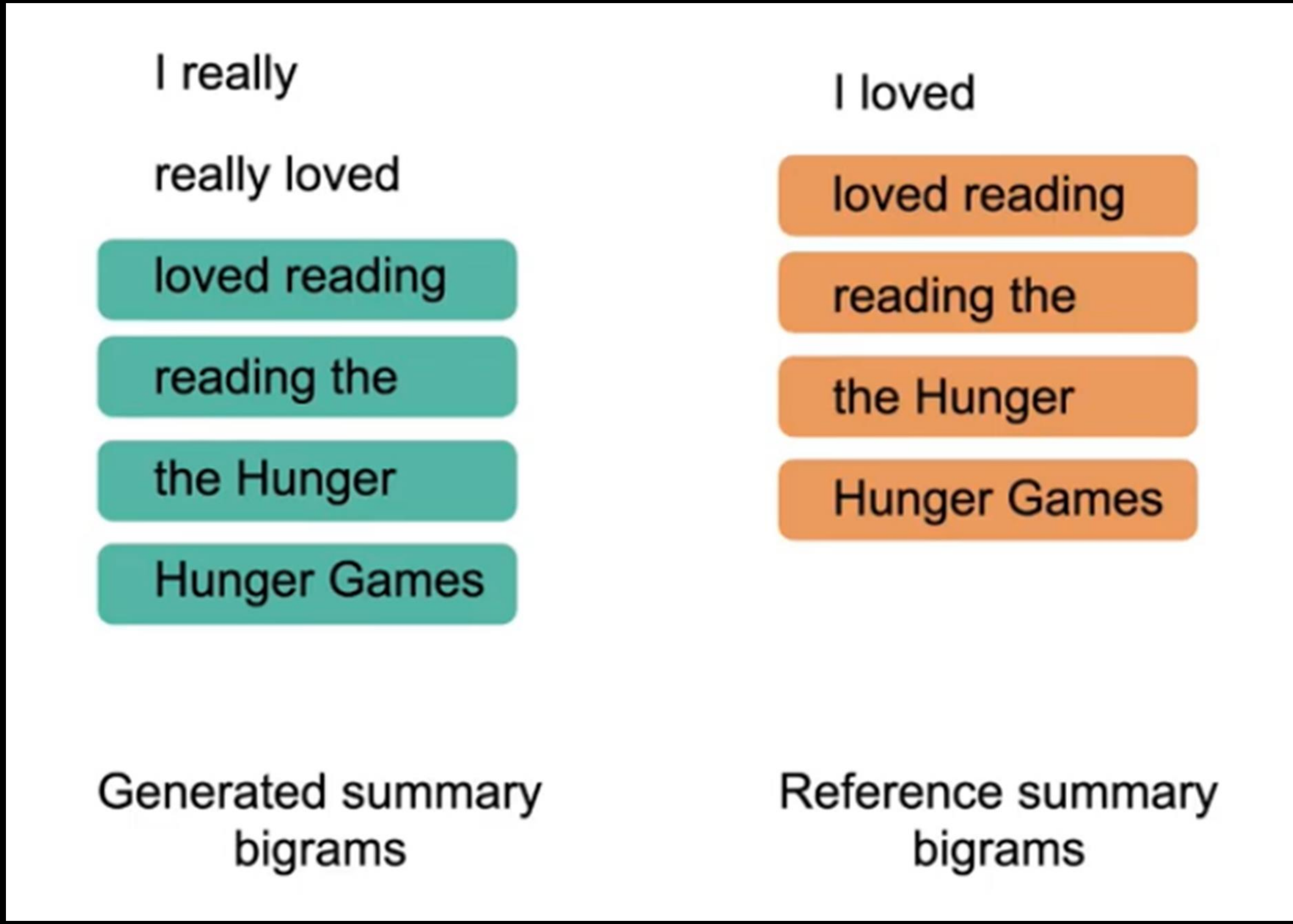
$$\text{ROUGE-1 recall} = \frac{\text{Num word matches}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-1 precision} = \frac{\text{Num word matches}}{\text{Num words in summary}} = \frac{6}{7}$$

$$\text{ROUGE-1 F1-score} = 2\left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)$$

HuggingFace. (2021, November 15). *What is the ROUGE metric?* https://www.youtube.com/watch?v=TMshhnrEXlg

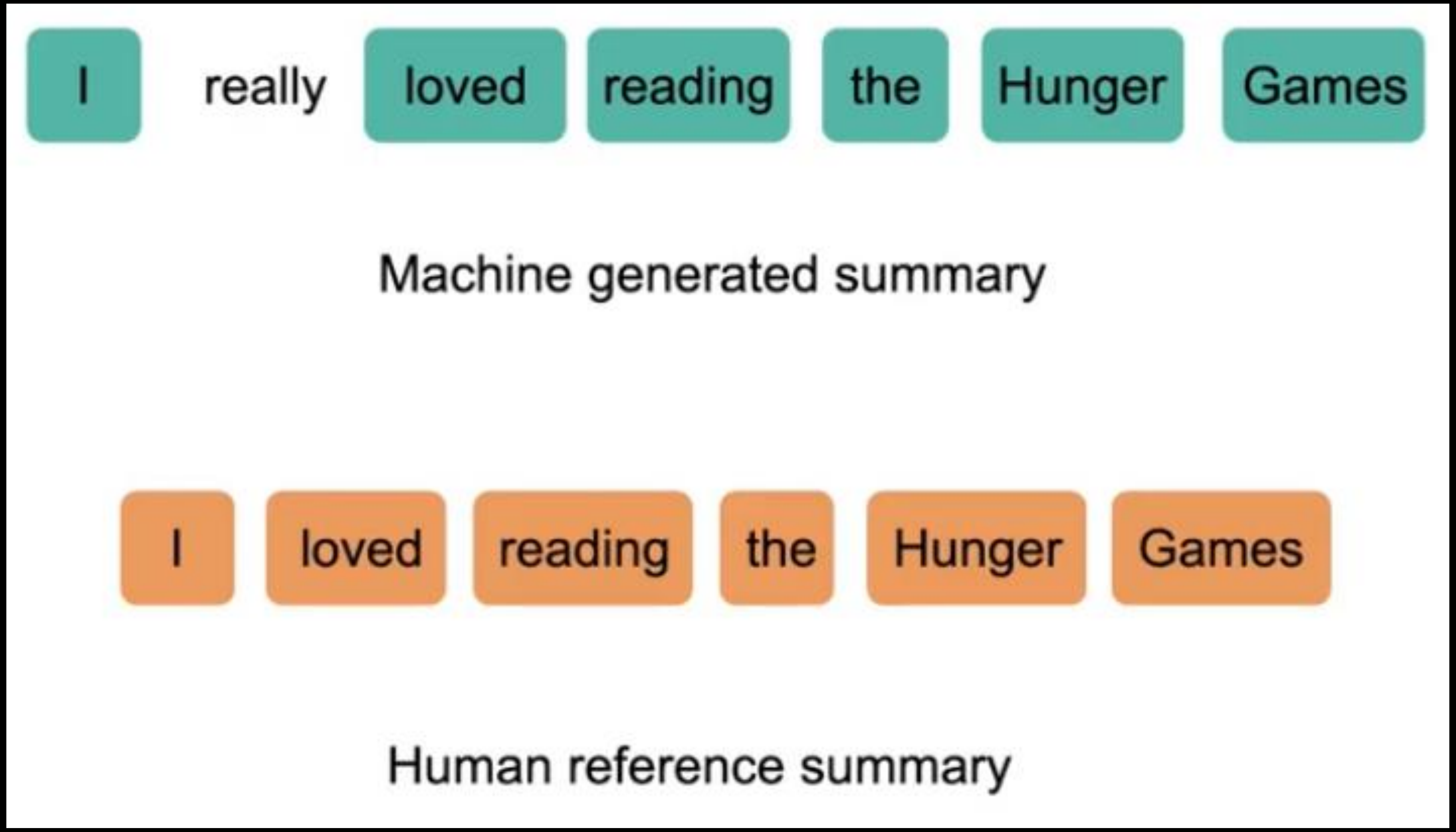
ROUGE-2 SCORE



$$\text{ROUGE-2 recall} = \frac{\text{Num bigram matches}}{\text{Num bigrams in reference}} = \frac{4}{5}$$

$$\text{ROUGE-2 precision} = \frac{\text{Num bigram matches}}{\text{Num bigram in summary}} = \frac{4}{6}$$

ROUGE-L SCORE



$$\text{ROUGE-L recall} = \frac{\text{LCS(gen, ref)}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-L precision} = \frac{\text{LCS(gen, ref)}}{\text{Num words in summary}} = \frac{6}{7}$$


What is the BLEU metric?
with Lewis

LIKELIHOOD OF A SEQUENCE

$$P(X) = \prod_{i=0}^{t} p(x_i \mid x_{<i})$$

Hugging Face is a startup based in New York City and Paris

p(word|context)

HuggingFace. (2021, November 15). *What is perplexity?* https://www.youtube.com/watch?v=NURcDHhYe98

CROSS-ENTROPY

$$CE(X) = -\frac{1}{t}\log P(X)$$

LOG-PERPLEXITY

$$PPL(X) = e^{CE(X)}$$
$$= e^{-\frac{1}{t}\sum_{i=0}^{t} \log p(x_i|x_{<i})}$$

Also see: https://towardsdatascience.com/perplexity-intuition-and-derivation-105dd481c8f3)

BENCHMARKS VS. DIRECT COMPARISONS

Spaces HuggingFaceH4/open_llm_leaderboard like 9.67k • Running on CPU UPGRADE

App Files Community 728

Open LLM Leaderboard

LLM Benchmark | Metrics through time | About | FAQ | Submit

Search models or licenses (e.g., 'model_name; license: MIT') and press ENTER...

Select columns to show

- [x] Average
- [x] ARC
- [x] HellaSwag
- [x] MMLU
- [x] TruthfulQA
- [x] Winogrande
- [x] GSM8K
- [ ] Type
- [ ] Architecture
- [ ] Precision
- [ ] Merged
- [ ] Hub License
- [ ] #Params (B)
- [ ] Hub ❤️
- [ ] Model sha

Hide models

- [x] Private or deleted
- [x] Contains a merge/moerge
- [x] Flagged
- [ ] MoE

Model types

- [x] 🟢 pretrained
- [x] 🟩 continuously pretrained
- [x] 🔶 fine-tuned on domain-specific datasets
- [x] 💬 chat models (RLHF, DPO, IFT, ...)
- [x] 🤝 base merges and moerges
- [x] ?

Precision

- [x] float16
- [x] bfloat16
- [x] 8bit
- [x] 4bit
- [x] GPTQ
- [x] ?

Model sizes (in billions of parameters)

- [x] ?
- [x] ~1.5
- [x] ~3
- [x] ~7
- [x] ~13
- [x] ~35
- [x] ~60
- [x] 70+

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande
🔶	CausalLM/34b-beta	73.04	70.56	84.2	85.6	58.38	81.29
🔶	NeverSleep/CausalLM-RP-34B	72.26	68	83.43	83.1	54.51	82.16
💬	MaziyarPanahi/Llama-3-70B-Instruct-DPO-v0.4	78.89	72.61	86.03	80.5	63.26	83.58
💬	MaziyarPanahi/Llama-3-70B-Instruct-DPO-v0.3	78.74	72.35	86	80.47	63.45	82.95
🟢	Qwen/Qwen1.5-110B	75.42	69.97	87.48	80.2	49.66	84.14
💬	MaziyarPanahi/Llama-3-70B-Instruct-DPO-v0.1	78.11	71.67	85.83	80.12	62.11	82.87
🔶	abhishek/autotrain-llama3-70b-orpo-v1	78.08	70.65	85.99	80.11	61.78	84.29
🔶	NeverSleep/Llama-3-Lumimaid-70B-v0.1	76.38	70.9	85.9	80.09	57.92	84.61
💬	abhishek/autotrain-llama3-70b-orpo-v2	78.17	70.9	86.09	80.07	62.82	84.93
💬	meta-llama/Meta-Llama-3-70B-Instruct	77.88	71.42	85.69	80.06	61.81	82.87

Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

Meta AI. (2024, April 18). Introducing Meta Llama 3: The most capable openly available LLM to date. Retrieved May 5, 2024, from Meta AI website: https://ai.meta.com/blog/meta-llama-3/

Astronomy

What is true for a type-Ia supernova?

A. This type occurs in binary systems.
B. This type occurs in young galaxies.
C. This type produces gamma-ray bursts.
D. This type produces high amounts of X-rays.
Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

A. directional selection.
B. stabilizing selection.
C. sexual selection.
D. disruptive selection
Answer: A

- **57 tasks**
- **Including elementary mathematics, US history, computer science, law, and more**
- **Models must possess extensive world knowledge and problem-solving ability**

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021, January 12). *Measuring Massive Multitask Language Understanding*. arXiv.

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **800,000** human pairwise comparisons to rank LLMs with the Bradley-Terry model and display the model ratings in Elo-scale. You can find more details in our paper.

Arena | Full Leaderboard

Total #models: **92**. Total #votes: **910,122**. Last updated: 2024-05-01.

📢 **NEW!** View leaderboard for different categories (e.g., coding, long user query)! This is still in preview and subject to change.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote 🗳️ at chat.lmsys.org!

***Rank (UB)**: model's ranking (upper-bound), defined by one + the number of models that are statistically better than the target model. Model A is statistically better than model B when A's lower-bound score is greater than B's upper-bound score (in 95% confidence interval). See Figure 3 below for visualization of the confidence intervals of model scores.

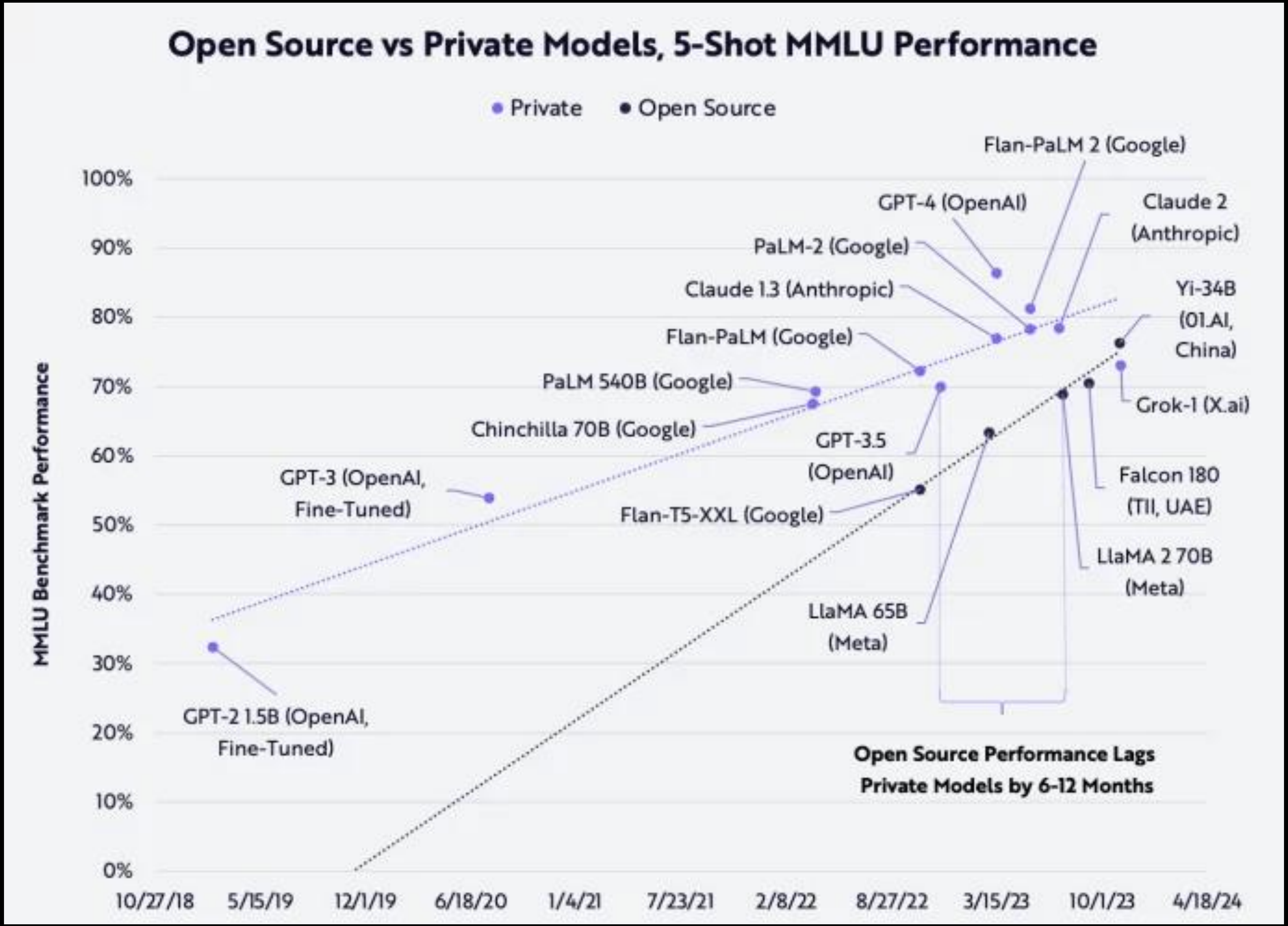
Category

Overall

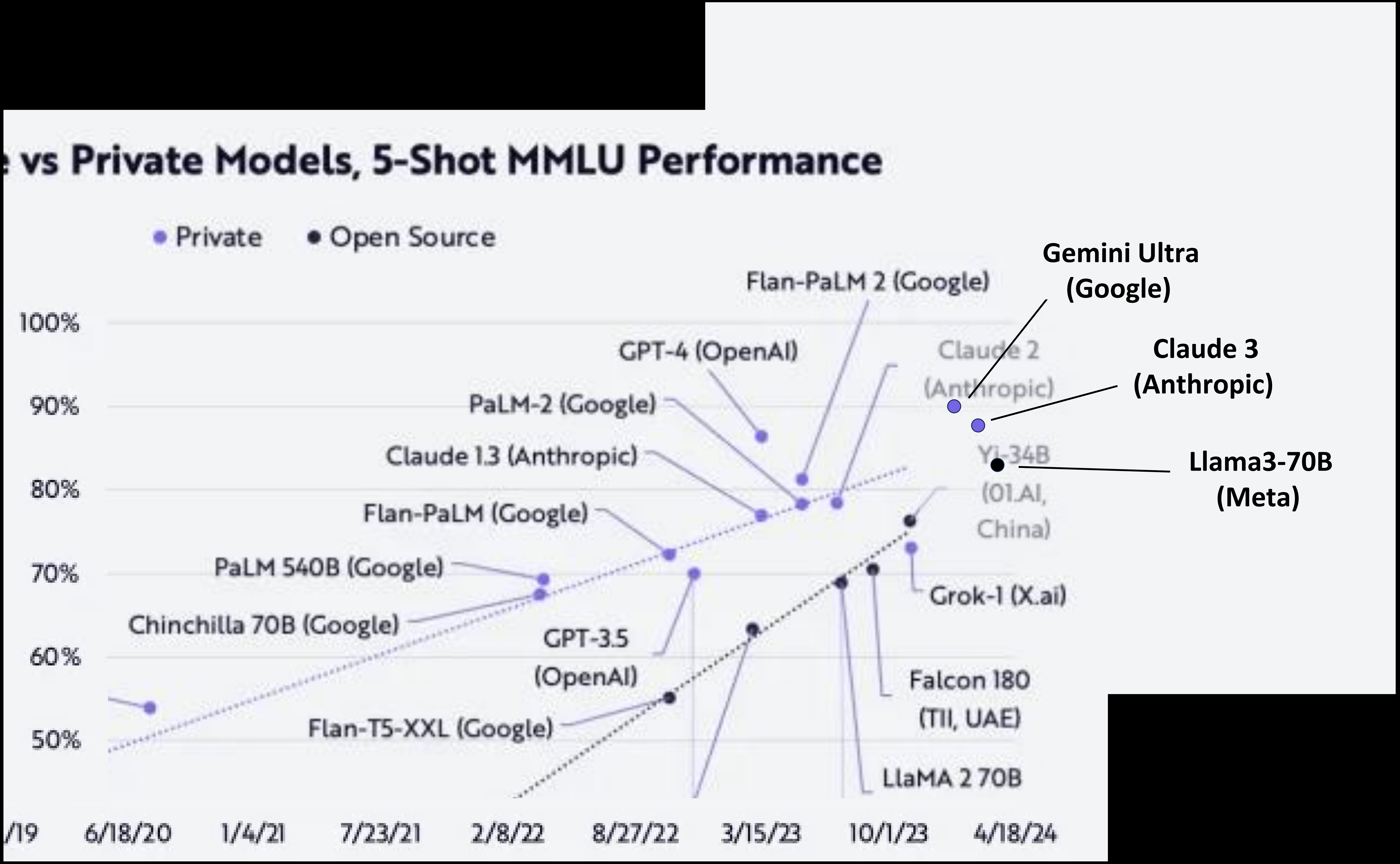
Overall Questions

#models: 92 (100%) **#votes: 910,122 (100%)**

Rank* (UB)	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI	Proprietary	2023/4
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	1250	+3/-3	39482	Google	Proprietary	2023/11
2	GPT-4-0125-preview	1247	+3/-2	67354	OpenAI	Proprietary	2023/12
6	Llama-3-70b-Instruct	1210	+3/-4	53404	Meta	Llama 3 Community	2023/12
6	Bard (Gemini Pro)	1209	+5/-6	12387	Google	Proprietary	Online
7	Claude 3 Sonnet	1201	+2/-3	78956	Anthropic	Proprietary	2023/8
9	Command R+	1191	+3/-3	44988	Cohere	CC-BY-NC-4.0	2024/3
9	GPT-4-0314	1190	+3/-4	52079	OpenAI	Proprietary	2021/9
11	Claude 3 Haiku	1181	+2/-3	69660	Anthropic	Proprietary	2023/8
12	GPT-4-0613	1165	+3/-3	70726	OpenAI	Proprietary	2021/9



ARK Invest. (2024). *BIG IDEAS 2024* [Annual Research Report]. ARK Investment Management LLC.



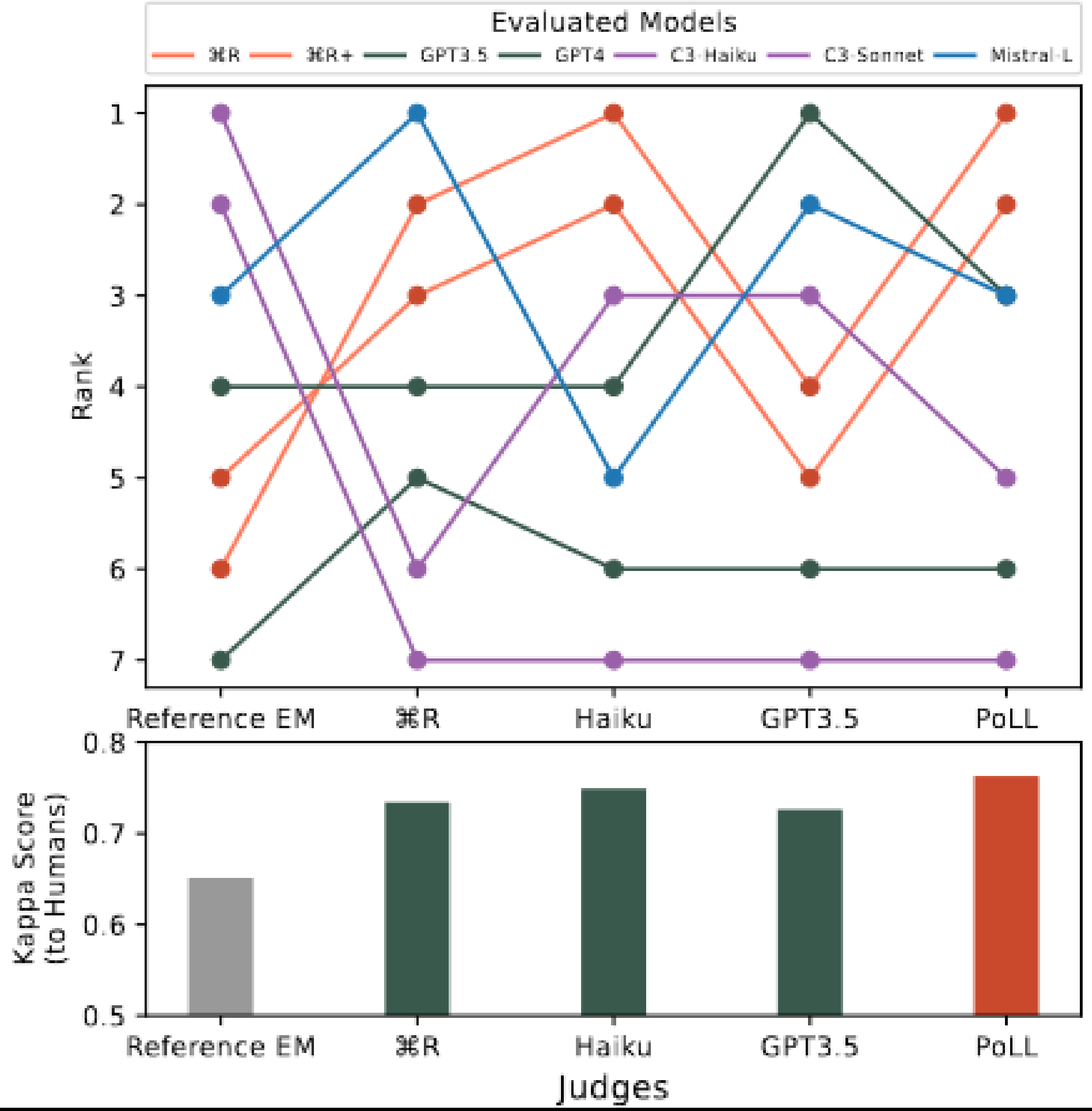
e vs Private Models, 5-Shot MMLU Performance
Private
Open Source
100%
90%
80%
70%
60%
50%
/19
6/18/20
1/4/21
7/23/21
2/8/22
8/27/22
3/15/23
10/1/23
4/18/24
Flan-PaLM 2 (Google)
GPT-4 (OpenAI)
PaLM-2 (Google)
Claude 1.3 (Anthropic)
Flan-PaLM (Google)
PaLM 540B (Google)
Chinchilla 70B (Google)
GPT-3.5
(OpenAI)
Flan-T5-XXL (Google)
Claude 2
(Anthropic)
Yi-34B
(01.AI,
China)
Grok-1 (X.ai)
Falcon 180
(TII, UAE)
LlaMA 2 70B
Gemini Ultra
(Google)
Claude 3
(Anthropic)
Llama3-70B
(Meta)

Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models

Pat Verga
Sebastian Hofstätter, Sophia Althammer, Yixuan Su
Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White
Patrick Lewis
Cohere

Abstract

As Large Language Models (LLMs) have become more advanced, they have outpaced our abilities to accurately evaluate their quality. Not only is finding data to adequately probe particular model properties difficult, but evaluating the correctness of a model's free-form generation alone is a challenge. To address this, many evaluations now rely on using LLMs themselves as judges to score the quality of outputs from other LLMs. Evaluations most commonly use a single large model like GPT-4. While this method has grown in popularity, it is costly, has been shown to introduce intra-model bias, and in this work, we find that very large mod-



Verga, P., Hofstatter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., … Lewis, P. (2024, May 1). *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*. arXiv.

A User-Centric Benchmark for Evaluating Large Language Models

Jiayin Wang[1], Fengran Mo[2], Weizhi Ma[3], Peijie Sun[1], Min Zhang[1*], Jian-Yun Nie[2]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China.
[2]Université de Montréal, Québec, Canada.
[3]Institute for AI Industry Research, Tsinghua University, Beijing, China.
JiayinWangTHU@gmail.com, z-m@tsinghua.edu.cn

Abstract

Large Language Models (LLMs) are essential tools to collaborate with users on different tasks. Evaluating their performance to serve users' needs in real-world scenarios is important. While many benchmarks have been created, they mainly focus on specific predefined model abilities. Few have covered the intended utilization of LLMs by real users. To address this oversight, we propose benchmarking LLMs from a user perspective in both dataset construction and evaluation designs. We first collect 1,846 real-world use cases with 15 LLMs from a user study with 712 participants from 23 countries. This forms the User

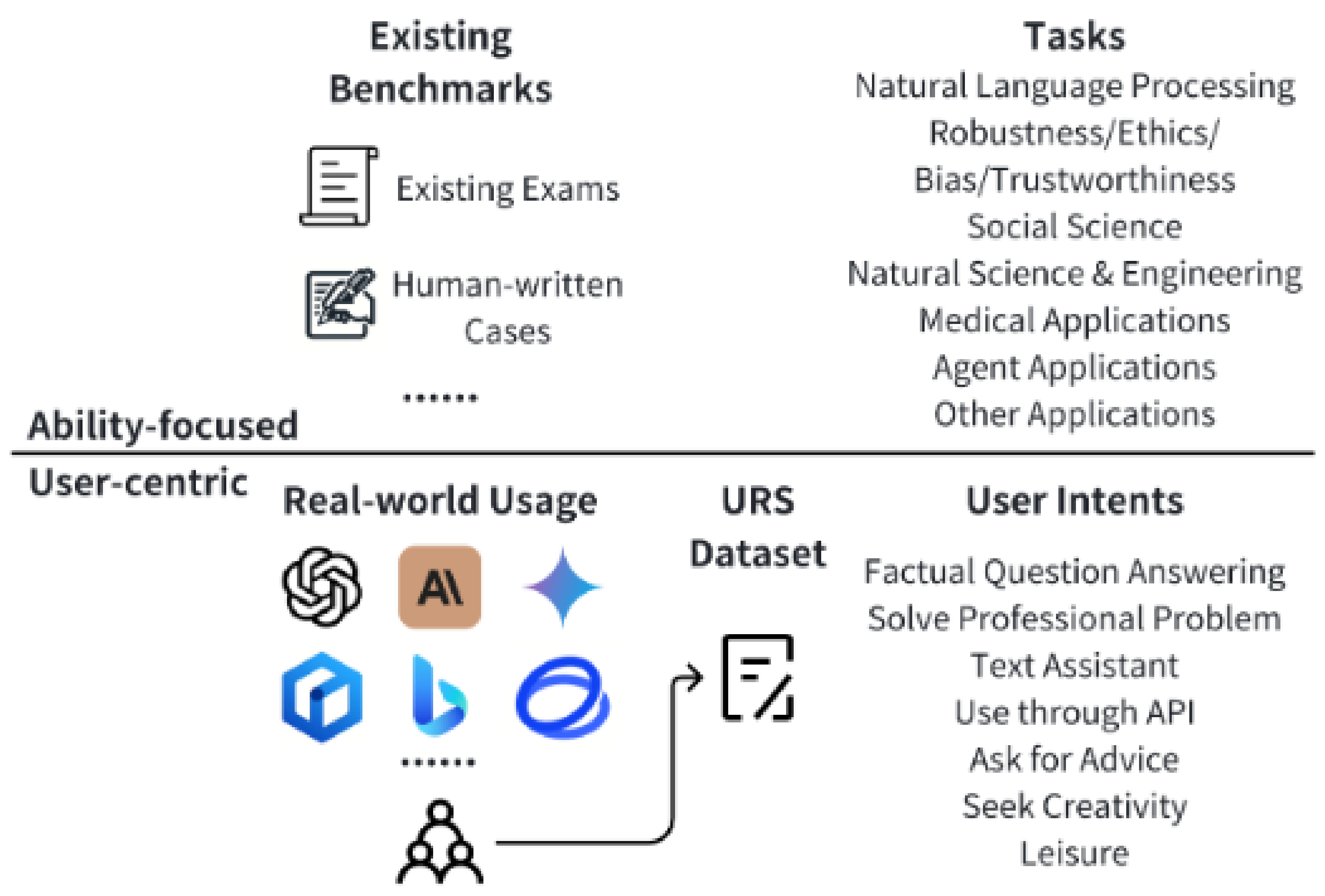


Figure 1: Existing benchmarks are mainly model ability-focused and categorized by tasks (Chang et al.,

Wang, J., Mo, F., Ma, W., Sun, P., Zhang, M., & Nie, J.-Y. (2024, April 22). *A User-Centric Benchmark for Evaluating Large Language Models*. arXiv.

APA PsycNet®

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal Information

Journal TOC

Search APA PsycNet

APA PsycArticles: Journal Article

An argument-based approach to validity.

© Request Permissions

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Outlines a general argument-based approach to validation, develops an interpretive argument for a placement test as an example, and examines some key properties in interpretive arguments. Validity is associated with the interpretation assigned to test scores rather than with the test scores or the test. The interpretation involves an argument leading from the scores to score-based statements or decisions, and the validity of the interpretation depends on the plausibility of this interpretive argument. The interpretive arguments associated with most test-score interpretations involve multiple inferences and assumptions. An explicit recognition of the inferences and assumptions in the interpretive argument makes it possible to identify the kinds of evidence needed to evaluate the argument. Evidence for the inferences and assumptions in the argument supports the interpretation, and evidence against any part of the argument casts doubt on the interpretation.

- **Validation focuses on evaluating the inferences that link the model results with their intended interpretations and uses.**
- **The Implications and associated decisions are most important for the validity of the results.**

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

PROJECT DISCUSSION

QUESTIONS

- Which approaches use similar projects?
- Which model do you want to fine-tune?
- How do you want to evaluate it?

- Web3 Coding Assistant — CodeLlama2, StarCoder // Julien, Kristian B., Anna-Valentina
- Socratic Assistant — Llama3 8B Chat // Ben, Julian
- Synthetic Data Generation for Event Data — Llama3 8B, GPT-3 .5 // Yorck, Kaan, Dikshyant, Khan
- Minimal Size Model for Conversations with Movie Characters — Phi2 // Christopher, Tural
- Training a Model for Diagnostics Based on Manuals — Llama3 8B // Christian W., Christian R., Dilip, James, Sina, Yildiz
- Financial Data Extraction — LeoLLM 7B // Nicolas
- Genome Chatbot — BioBERT? // Muhammad
- Small Size Language Learning Assistant — Phi3 Mini, LeoLLM, Sauerkraut// Rafael, Ilhay, Philip
- Small, open-source, multilingual function-calling agents — Phi3 Mini, RWKI, Tiny Llama // Jeremy, Boran

TASKS UNTIL NEXT WEEK

- Decide on a baseline model and implement an evaluation chain for your base model.