CS11-711 Advanced NLP

Tour of Modern LLMs (and surrounding topics)

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Site
https://phontron.com/class/anlp2024/

What Makes a Model?

- Architecture decisions
- Data decisions
- Training decisions

Open vs. Closed Access

Open/Closed Access
(e.g. Liang et al. 2022)

- **Weights:** open? described? closed?
- **Inference Code:** open? described? closed?
- **Training Code:** open? described? closed?
- **Data:** open? described? closed?

Licenses and Permissiveness

- **Public domain, CC-0:** old copyrighted works and products of US government workers
- **MIT, BSD:** very few restrictions
- **Apache, CC-BY:** must acknowledge owner
- **GPL, CC-BY-SA:** must acknowledge and use same license for derivative works
- **CC-NC:** cannot use for commercial purposes
- **LLaMa, OPEN-RAIL:** various other restrictions
- **No License:** all rights reserved, but can use under fair use

Fair Use

- US **fair use** doctrine — can use copyrighted material in some cases
- A gross simplification:
  - **Quoting** a small amount of material → likely OK
  - **Doesn't diminish** commercial value → possibly OK
  - Use for **non-commercial** purposes → possibly OK
- Most data on the internet is copyrighted, so model training is currently done assuming fair use
- But there are lawsuits!

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

Why Restrict Model Access?

- **Commercial Concerns:** Want to make money from the models
- **Safety:** Limited release prevents possible misuse
- **Legal Liability:** Training models on copyrighted data is a legal/ethical gray area

English-Centric Open Models

Birds-eye View

- Open source/reproducible:
  - **Pythia:** Fully open, many sizes/checkpoints
  - **OLMo:** Possibly strongest reproducible model
- Open weights:
  - **LLaMa2:** Most popular, heavily safety tuned
  - **Mistral/Mixtral:** Strong and fast model, several European languages
  - **Qwen:** Strong, more multilingual - particularly en/zh

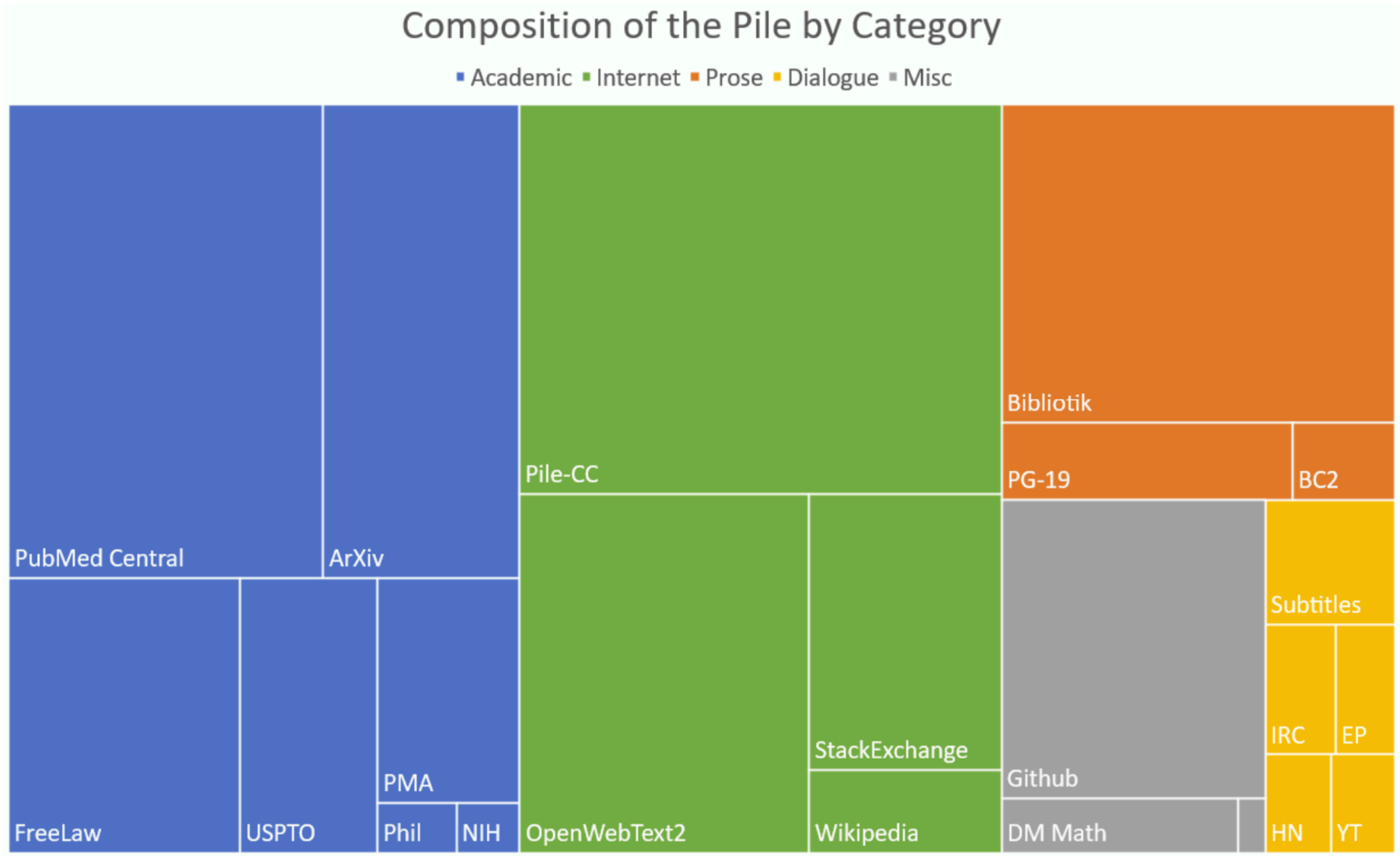
Pythia - Overview

- **Creator:** EleutherAI
- **Goal:** Joint understanding of model training dynamics and scaling
- **Unique features:** 8 model sizes 70M-12B, 154 checkpoints for each

Arch	Transformer+RoPE+SwiGLU, context 2k (cf LLaMa 4k), parametric LN
Data	Trained on 300B tokens of The Pile (next slide), or deduped 207B
Train	LR scaled inversely to model size (7B=1.2e-4), batch size 2M tokens

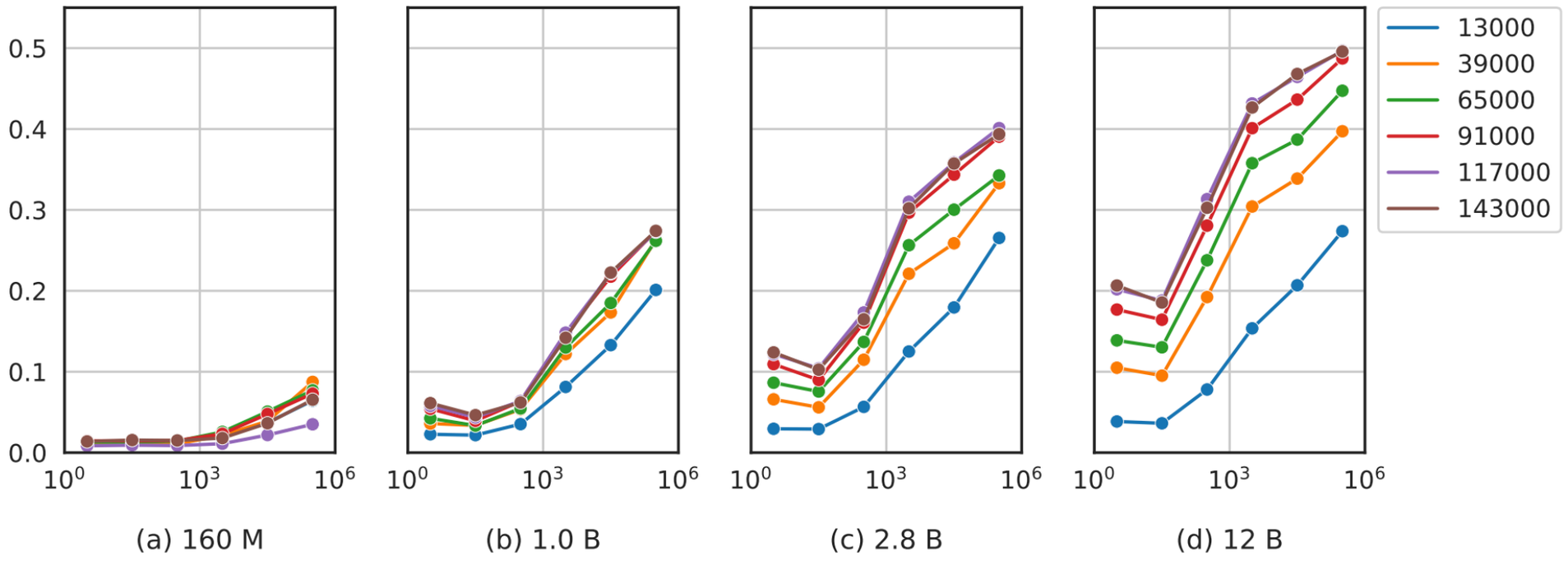
The Pile

- A now-standard 800GB dataset of lots of text/code



Pythia - Findings

- Some insights into training dynamics, e.g. larger models memorize facts more quickly (x axis: fact frequency, legend: training step)



(a) 160 M (b) 1.0 B (c) 2.8 B (d) 12 B

- It is possible to intervene on data to reduce gender bias

OLMo - Overview

- **Creator:** AI2 Allen Institute for AI
- **Goal:** Better science of state-of-the-art LMs
- **Unique features:** Top performance of fully documented model, instruction tuned etc.

Arch	Transformer+RoPE+SwiGLU, context 4k, non-parametric LN
Data	Trained on 2.46T tokens of Dolma corpus (next slide)
Train	LR scaled inversely to model size (7B=3e-4), batch size 4M tokens

Dolma

- 3T token corpus created and released by AI2 for LM training
- a pipeline of (1) language filtering, (2) quality filtering, (3) content filtering, (4) deduplication, (5) multi-source mixing, and (6) tokenization

Source	**Doc Type**	**UTF-8 bytes** *(GB)*	**Documents** *(millions)*	**Unicode words** *(billions)*	**Llama tokens** *(billions)*
Common Crawl	web pages	9,022	3,370	1,775	2,281
The Stack	code	1,043	210	260	411
C4	web pages	790	364	153	198
Reddit	social media	339	377	72	89
PeS2o	STEM papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7	4.3
Total		**11,519**	**4,367**	**2,318**	**3,059**

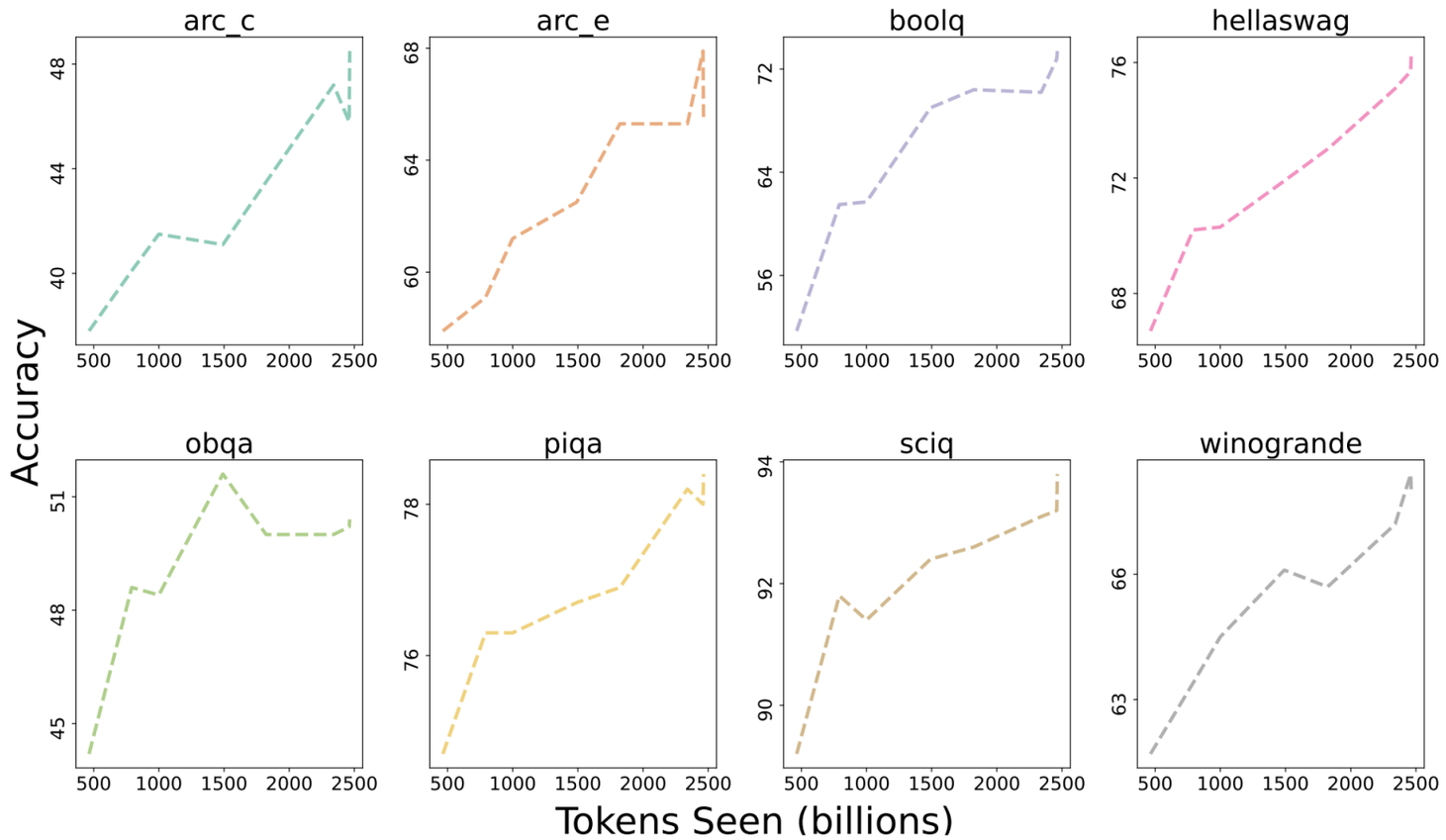
OLMo - Findings

- Competitive average performance

7B Models	arc challenge	arc easy	boolq	hella-swag	open bookqa	piqa	sciq	wino-grande	avg.
Falcon	47.5	70.4	74.6	75.9	53.0	78.5	93.9	68.9	70.3
LLaMA	44.5	67.9	75.4	76.2	51.2	77.2	93.9	70.5	69.6
Llama 2	48.5	69.5	80.2	76.8	48.4	76.7	94.5	69.4	70.5
MPT	46.5	70.5	74.2	77.6	48.6	77.3	93.7	69.9	69.8
Pythia	44.1	61.9	61.1	63.8	45.0	75.1	91.1	62.0	63.0
RPJ-INCITE	42.8	68.4	68.6	70.3	49.4	76.0	92.9	64.7	66.6
OLMo-7B	48.5	65.4	73.4	76.4	50.4	78.4	93.8	67.9	69.3

Table 6: Zero-shot evaluation of OLMo-7B and 6 other publicly available comparable model checkpoints on 8 core tasks from the downstream evaluation suite described in Section 2.4. For OLMo-7B, we report results for the 2.46T token checkpoint.

- Performance increases constantly w/ training



LLaMa2 - Overview

- **Creator:** Meta
- **Goal:** Strong and safe open LM w/ base+chat versions
- **Unique features:** Open model with strong safeguards and chat tuning, good performance

Arch	Transformer+RoPE+SwiGLU, context 4k, RMSNorm
Data	Trained on “public sources, up-sampling the most factual sources”, LLaMa 1 has more info (next page), total 2T tokens
Train	7B=3e-4, batch size 4M tokens

LLaMa 1 - Training Data

- Several sources, with more reliable source upsampled

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

LLaMa2 - Reward Model

- LLaMa 2 dev put a large emphasis on safety
- **Step 1:** Collect data for reward modeling

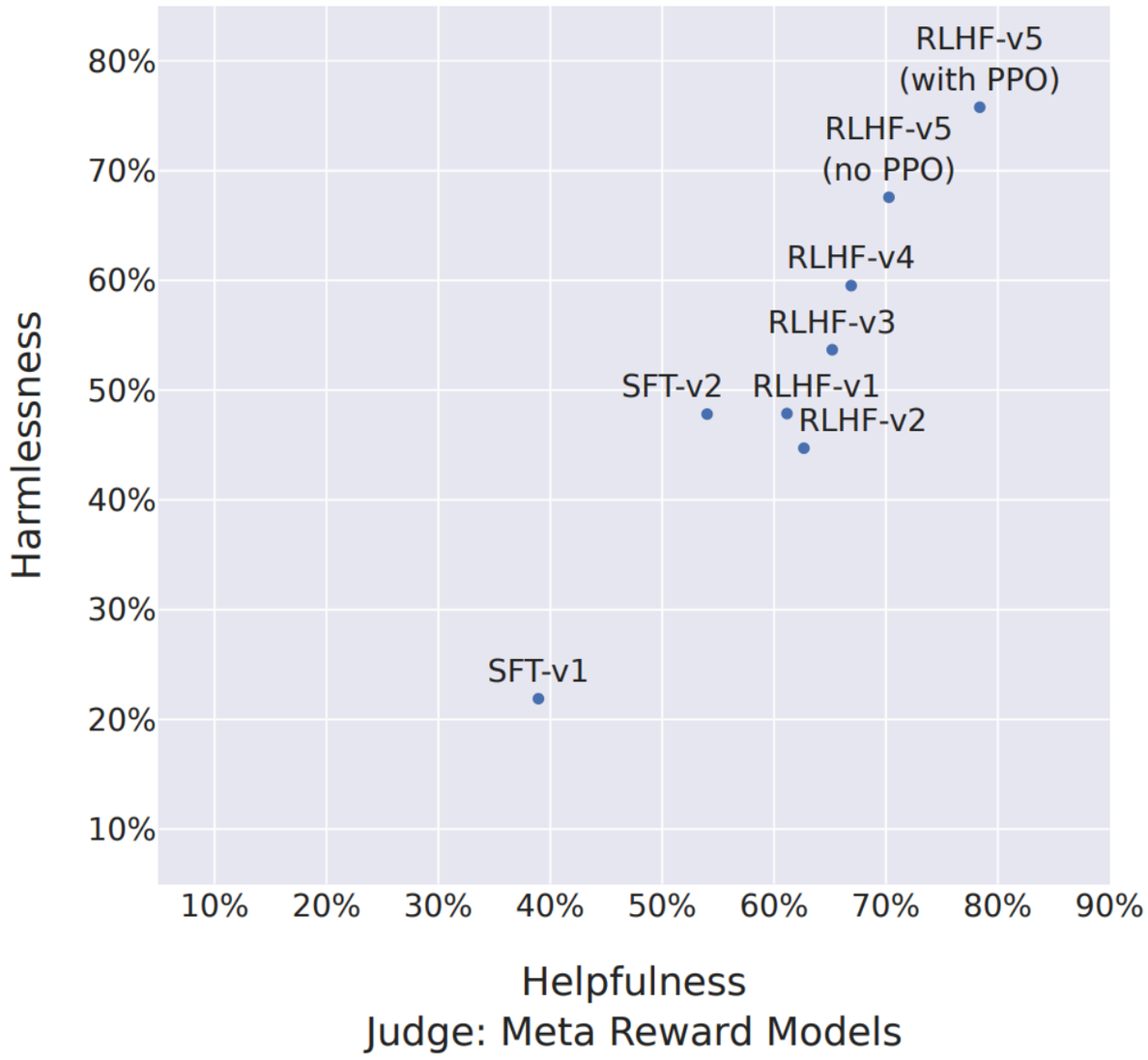
Dataset	Num. of Comparisons	Avg. # Turns per Dialogue	Avg. # Tokens per Example	Avg. # Tokens in Prompt	Avg. # Tokens in Response
Anthropic Helpful	122,387	3.0	251.5	17.7	88.4
Anthropic Harmless	43,966	3.0	152.5	15.7	46.4
OpenAI Summarize	176,625	1.0	371.1	336.0	35.1
OpenAI WebGPT	13,333	1.0	237.2	48.3	188.9
StackExchange	1,038,480	1.0	440.2	200.1	240.2
Stanford SHP	74,882	1.0	338.3	199.5	138.8
Synthetic GPT-J	33,139	1.0	123.3	13.0	110.3
Meta (Safety & Helpfulness)	1,418,091	3.9	798.5	31.4	234.1
Total	2,919,326	1.6	595.7	108.2	216.9

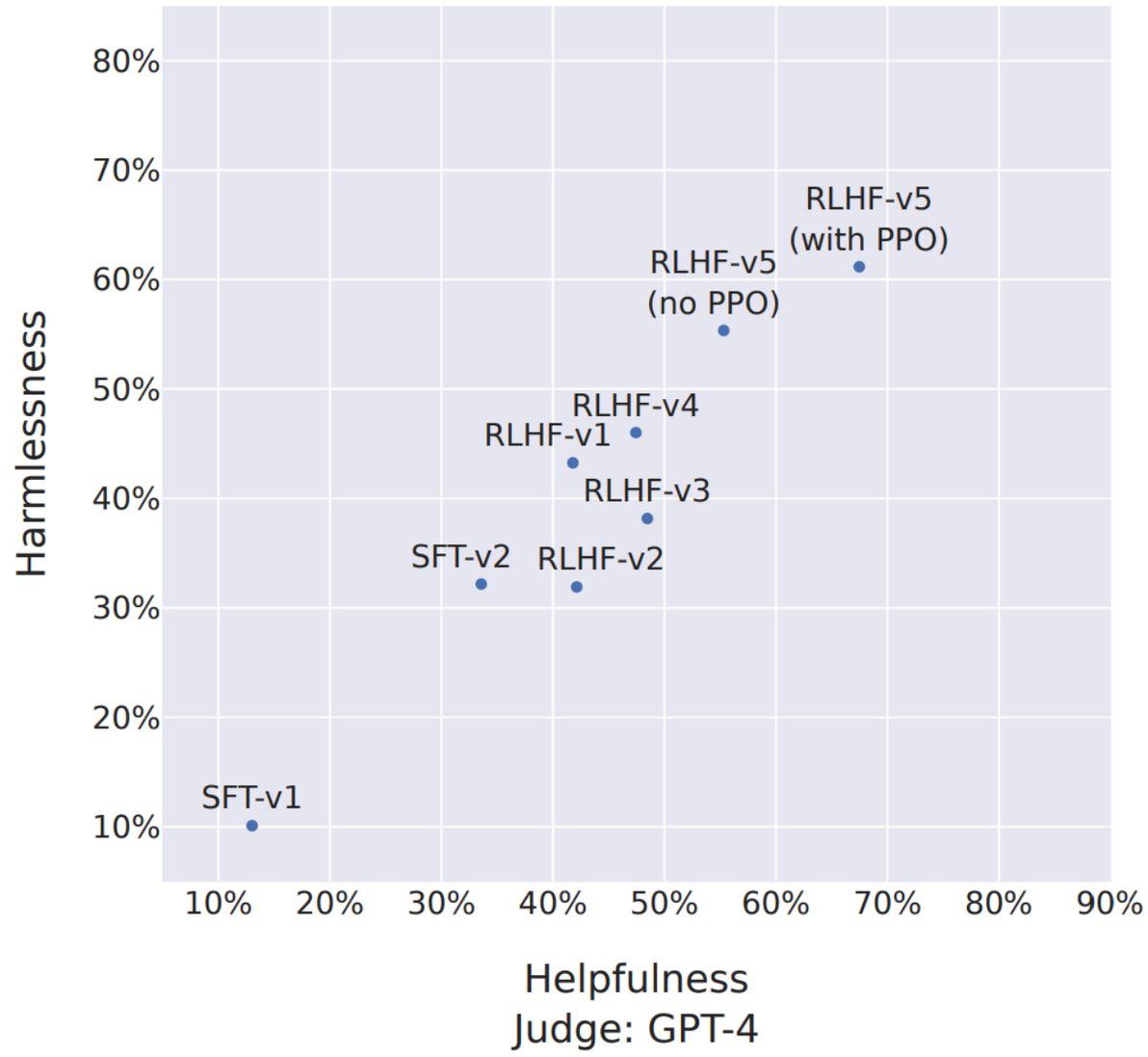
- **Step 2:** Train model to follow these preferences

	Meta Helpful.	Meta Safety	Anthropic Helpful	Anthropic Harmless	OpenAI Summ.	Stanford SHP	Avg
SteamSHP-XL	52.8	43.8	66.8	34.2	54.7	75.7	55.3
Open Assistant	53.8	53.4	67.7	68.4	71.7	55.0	63.0
GPT4	58.6	58.1	-	-	-	-	-
Safety RM	56.2	64.5	55.4	74.7	71.7	65.2	64.3
Helpfulness RM	63.2	62.8	72.0	71.0	75.5	80.0	70.6

LLaMa2 - RLHF

- Train model using reward model





- Each round of RLHF improves final model

LLaMa2 - Chat Instruction Following

- Prompting to follow instructions through “context distillation” (Askell et al. 2021) or “ghost attention”

Data Generation Phase

```
System: Write in only emojis.
User: Write in only emojis. Say hello.
Assistant: [generates]👋
User: Write in only emojis. How are you doing.
Assistant: [generates]😊💗
```

Training Phase

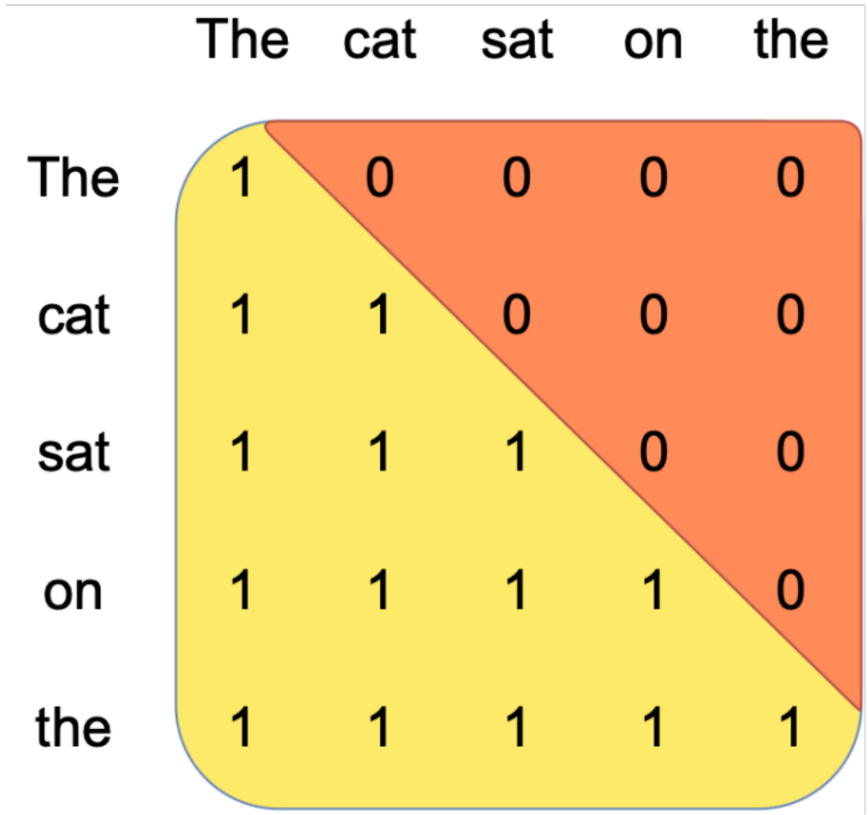
```
System: Write in only emojis.
User: Say hello.
Assistant: 👋
User: How are you doing.
Assistant: 😊💗
```

Mistral/Mixtral - Overview

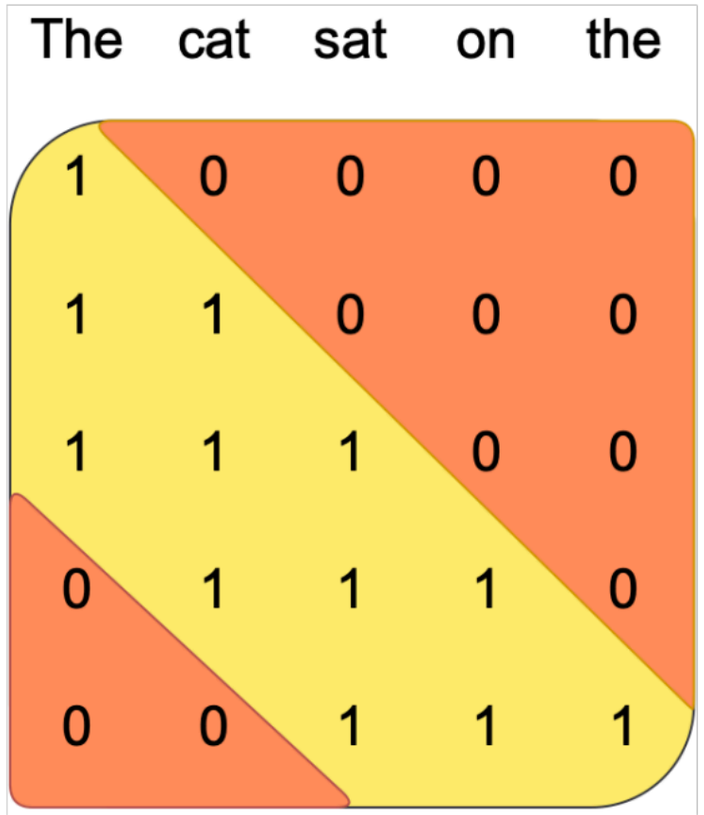
- **Creator:** MISTRAL AI_
- **Goal:** Strong and somewhat multilingual open LM
- **Unique features:** Speed optimizations, including GQA and Mixture of Experts

Arch	Transformer+RoPE+SwiGLU, context 4k, RMSNorm, sliding window attention. Mixtral has 8x experts in feed-forward layer
Data	Not disclosed? But includes English and European languages
Train	Not disclosed?

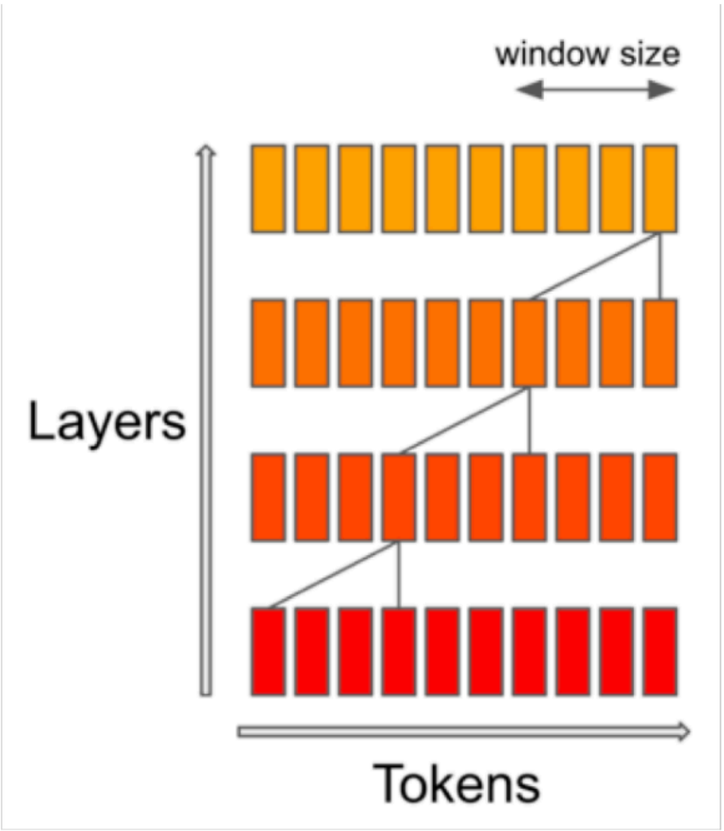
Mistral - Sliding Window Attention



Vanilla Attention



Sliding Window Attention



Effective Context Length

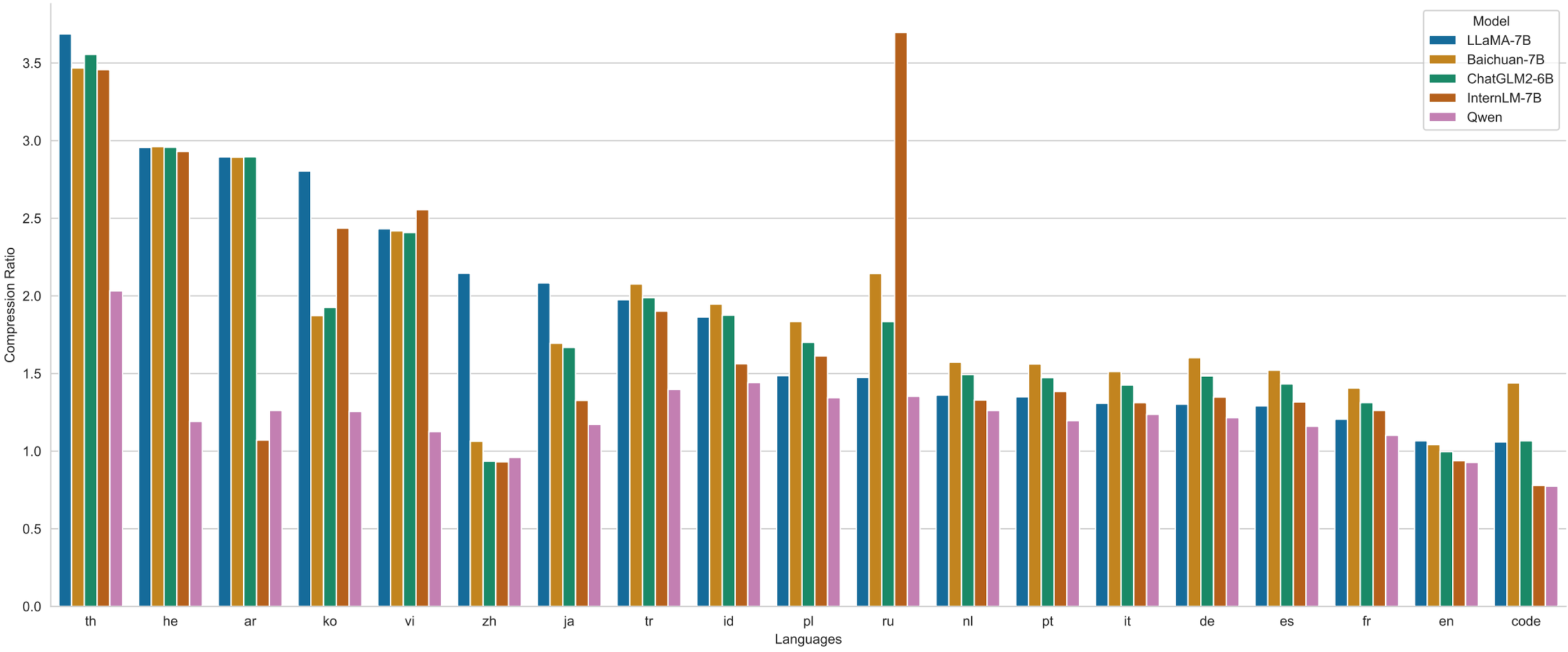
Qwen - Overview

- **Creator:** Alibaba
- **Goal:** Strong multilingual (esp. English and Chinese) LM
- **Unique features:** Large vocabulary for multilingual support, strong performance

Arch	Transformer+RoPE+SwiGLU, context 4k, RMSNorm, bias in attention layer
Data	Trained on multilingual data + instruction data at pre-training time, 2-3T tokens
Train	3e-4, batch size 4M tokens

Qwen - Multilinguality

- Token compression ratio re: XLM-R (lower is better)



Other Models

Code Models

- **StarCoder 2** — by Big Science (leads: Hugging Face + Service Now), fully open model
- **CodeLlama** — by Meta, code adaptation of LLaMa
- **DeepSeek Coder** — by DeepSeek, strong performance across many tasks
- More in code generation class!

Math Models

- **LLeMa** — by EleutherAI and others, model for math theorem proving trained on proof pile
- **DeepSeek Math** — by DeepSeek, finds math-related pages on the web

Scince Model: Galactica

- Model for science trained by Meta
- Diverse set of interesting training data

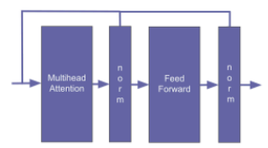
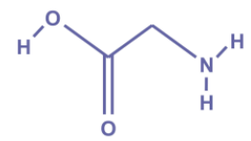
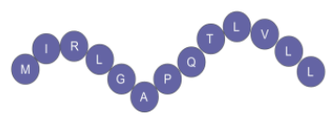
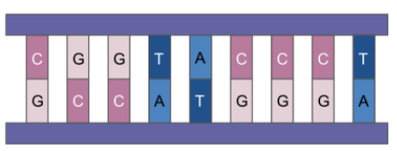
Modality	Entity	Sequence	
Text	Abell 370	`Abell 370 is a cluster...`	
LaTeX	Schwarzschild radius	`r_{s} = \frac{2GM}{c^2}`	$r_s = \frac{2GM}{c^2}$
Code	Transformer	`class Transformer(nn.Module)`	
SMILES	Glycine	`C(C(=O)O)N`	
AA Sequence	Collagen α-1(II) chain	`MIRLGAPQTL..`	
DNA Sequence	Human genome	`CGGTACCCTC..`	

Table 1: Tokenizing Nature. Galactica trains on text sequences that represent scientific phenomena.

Closed Models

GPT-4 - Overview

- **Creator:** OpenAI
- De-facto standard “strong” language model
- Tuned to be good as a chat-based assistant
- Accepts image inputs
- Supports calling external tools through “function calling” interface

Gemini

- **Creator:** Google DeepMind
- Performance competitive with corresponding GPT models (Gemini Pro 1.0 ~ gpt-3.5, Gemini Ultra 1.0 ~ gpt-4)
- Pro 1.5 supports very long inputs, 1-10M tokens
- Supports image and video inputs
- Can generate images natively

Claude 3 - Overview

- **Creator:** ANTHROP\C
- Context window up to 200k
- Allows for processing images
- Overall strong results competitive with GPT-4

Questions?