



The Llama 3 Herd of Models

Govind Ramesh and Zheng Wang

Georgia Tech, Atlanta

Evaluations:

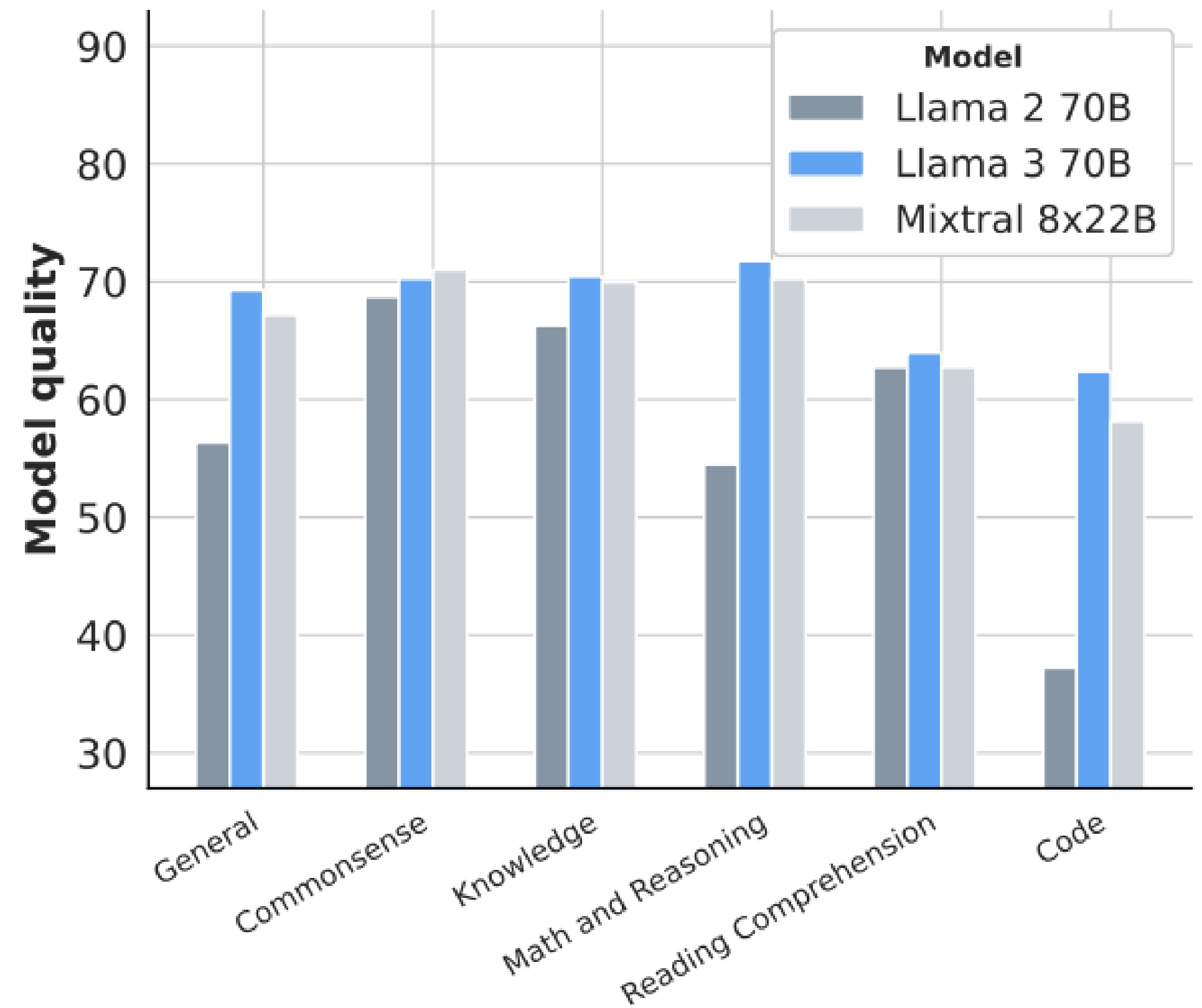
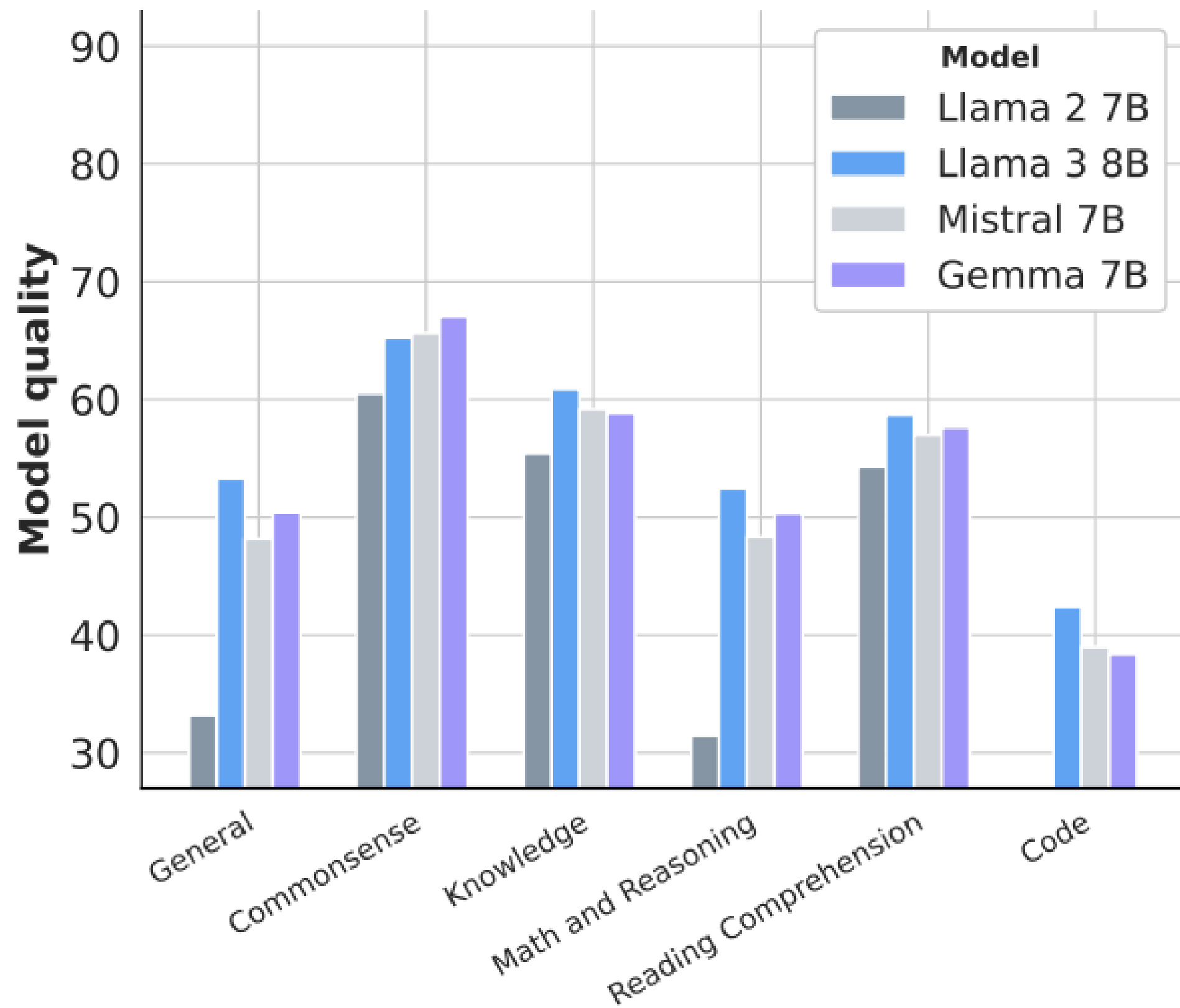
- Pre-trained model
- Post-trained model
- Safety characteristics

5. Pre-trained Models

Reading Comprehension	SQuAD V2 (Rajpurkar et al., 2018), QuaC (Choi et al., 2018), RACE (Lai et al., 2017),
Code	HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021),
Commonsense reasoning/understanding	CommonSenseQA (Talmor et al., 2019), PiQA (Bisk et al., 2020), SiQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), WinoGrande (Sakaguchi et al., 2021)
Math, reasoning, and problem solving	GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021b), ARC Challenge (Clark et al., 2018), DROP (Dua et al., 2019), WorldSense (Benchekroun et al., 2023)
Adversarial	Adv SQuAD (Jia and Liang, 2017), Dynabench SQuAD (Kiela et al., 2021), GSM-Plus (Li et al., 2024c) PAWS (Zhang et al., 2019)
Long context	QuALITY (Pang et al., 2022), many-shot GSM8K (An et al., 2023a)
Aggregate	MMLU (Hendrycks et al., 2021a), MMLU-Pro (Wang et al., 2024b), AGIEval (Zhong et al., 2023), BIG-Bench Hard (Suzgun et al., 2023)

Table 8 Pre-training benchmarks by category. Overview of all benchmarks we use to evaluate pre-trained Llama 3 models, grouped by capability category.

5. Pre-trained Models



5. Pre-trained Models

	Reading Comprehension		
	SQuAD	QuAC	RACE
Llama 3 8B	77.0 \pm 0.8	44.9 \pm 1.1	54.3 \pm 1.4
Mistral 7B	73.2 \pm 0.8	44.7 \pm 1.1	53.0 \pm 1.4
Gemma 7B	81.8 \pm 0.7	42.4 \pm 1.1	48.8 \pm 1.4
Llama 3 70B	81.8 \pm 0.7	51.1 \pm 1.1	59.0 \pm 1.4
Mixtral 8 \times 22B	84.1 \pm 0.7	44.9 \pm 1.1	59.2 \pm 1.4
Llama 3 405B	81.8 \pm 0.7	53.6 \pm 1.1	58.1 \pm 1.4
GPT-4	–	–	–
Nemotron 4 340B	–	–	–
Gemini Ultra	–	–	–

	Code	
	HumanEval	MBPP
Llama 3 8B	37.2 \pm 7.4	47.6 \pm 4.4
Mistral 7B	30.5 \pm 7.0	47.5 \pm 4.4
Gemma 7B	32.3 \pm 7.2	44.4 \pm 4.4
Llama 3 70B	58.5 \pm 7.5	66.2 \pm 4.1
Mixtral 8 \times 22B	45.1 \pm 7.6	71.2 \pm 4.0
Llama 3 405B	61.0 \pm 7.5	73.4 \pm 3.9
GPT-4	67.0 \pm 7.2	–
Nemotron 4 340B	57.3 \pm 7.6	–
Gemini Ultra	74.4 \pm 6.7	–

5. Pre-trained Models

Robustness

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 5: A question from the Professional Medicine task.

MMLU
Dataset

Performance can be sensitive to arbitrary changes in problem setup.

- few-shot label bias
- label variants
- answer order
- prompt format

5. Pre-trained Models

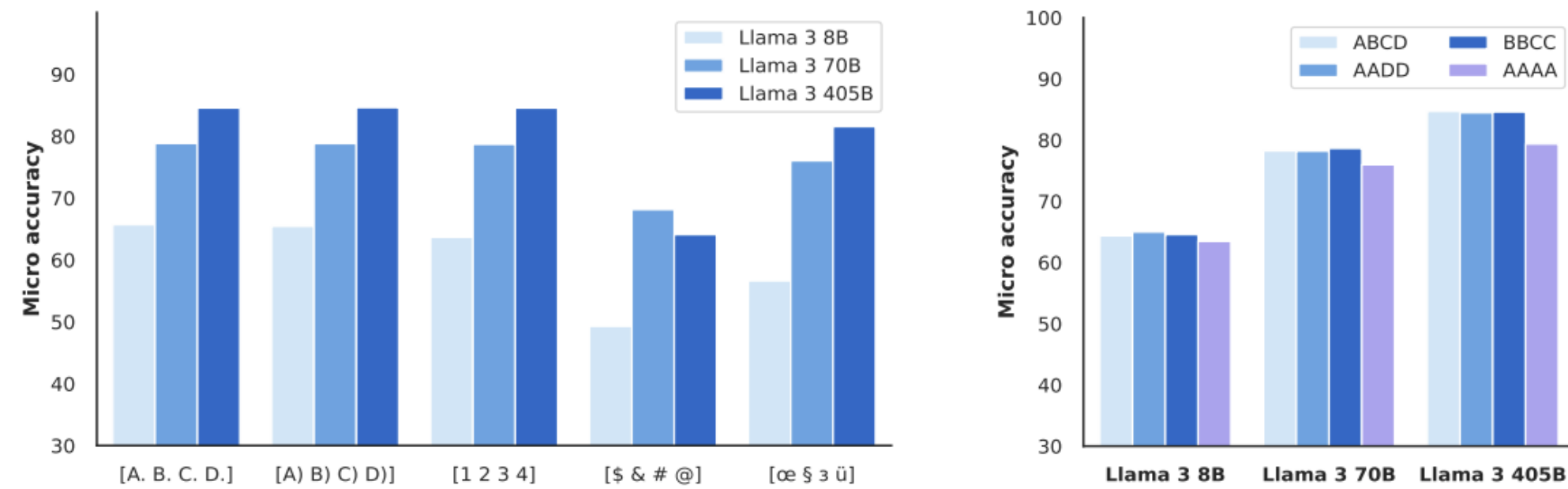


Figure 13 Robustness of our pre-trained language models to different design choices in the MMLU benchmark. *Left:* Performance for different label variants. *Right:* Performance for different labels present in few-shot examples.

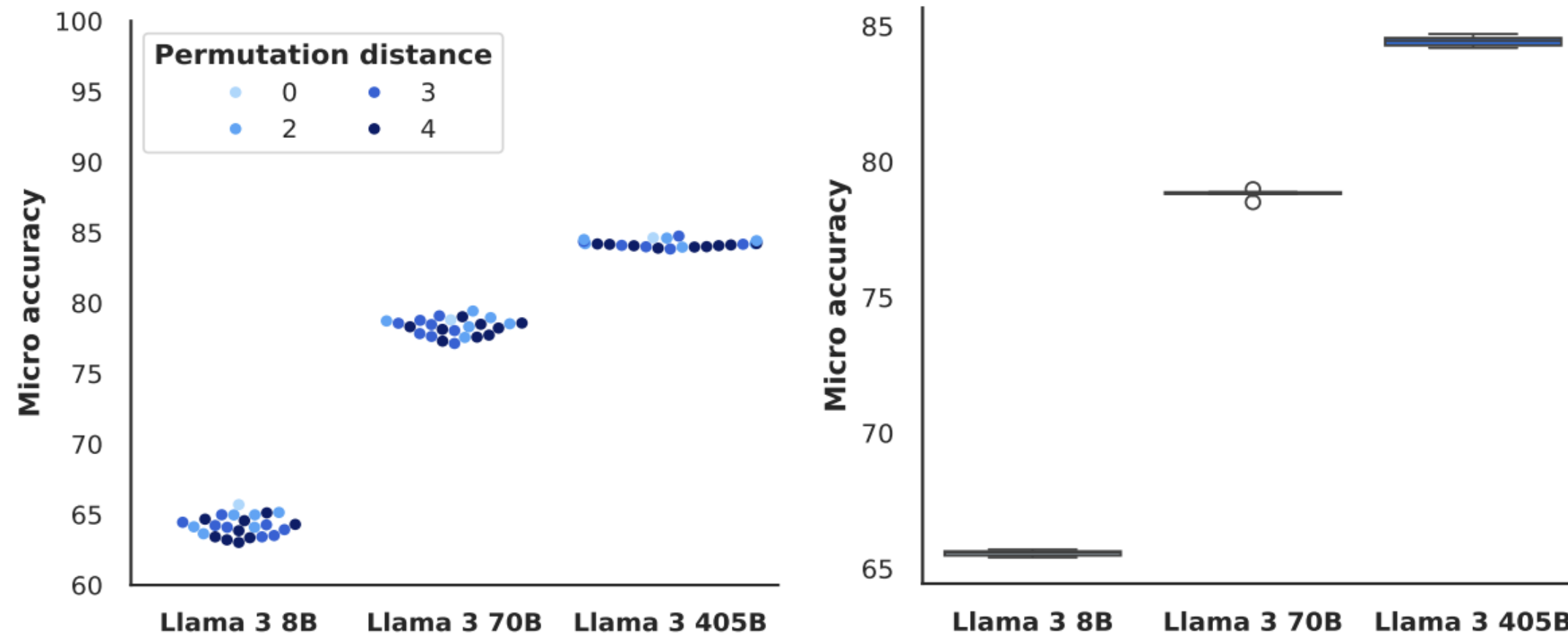


Figure 14 Robustness of our pre-trained language models to different design choices in the MMLU benchmark. *Left:* Performance for different answer orders. *Right:* Performance for different prompt formats.

Adversarial Benchmarks

Tests performance on tasks designed to be challenging

- question answering - Adversarial SQuAD and Dynabench SQuAD
- mathematical reasoning - GSM-Plus
 - A robe takes 2 bolts of blue fiber and half that much white fiber. If each bolt of fiber costs \$5 but there's a special discount today that reduces the price of each bolt by \$2, how many bolts in total does it take to make the robe?
- paraphrase detection - PAWS

Article: Super Bowl 50

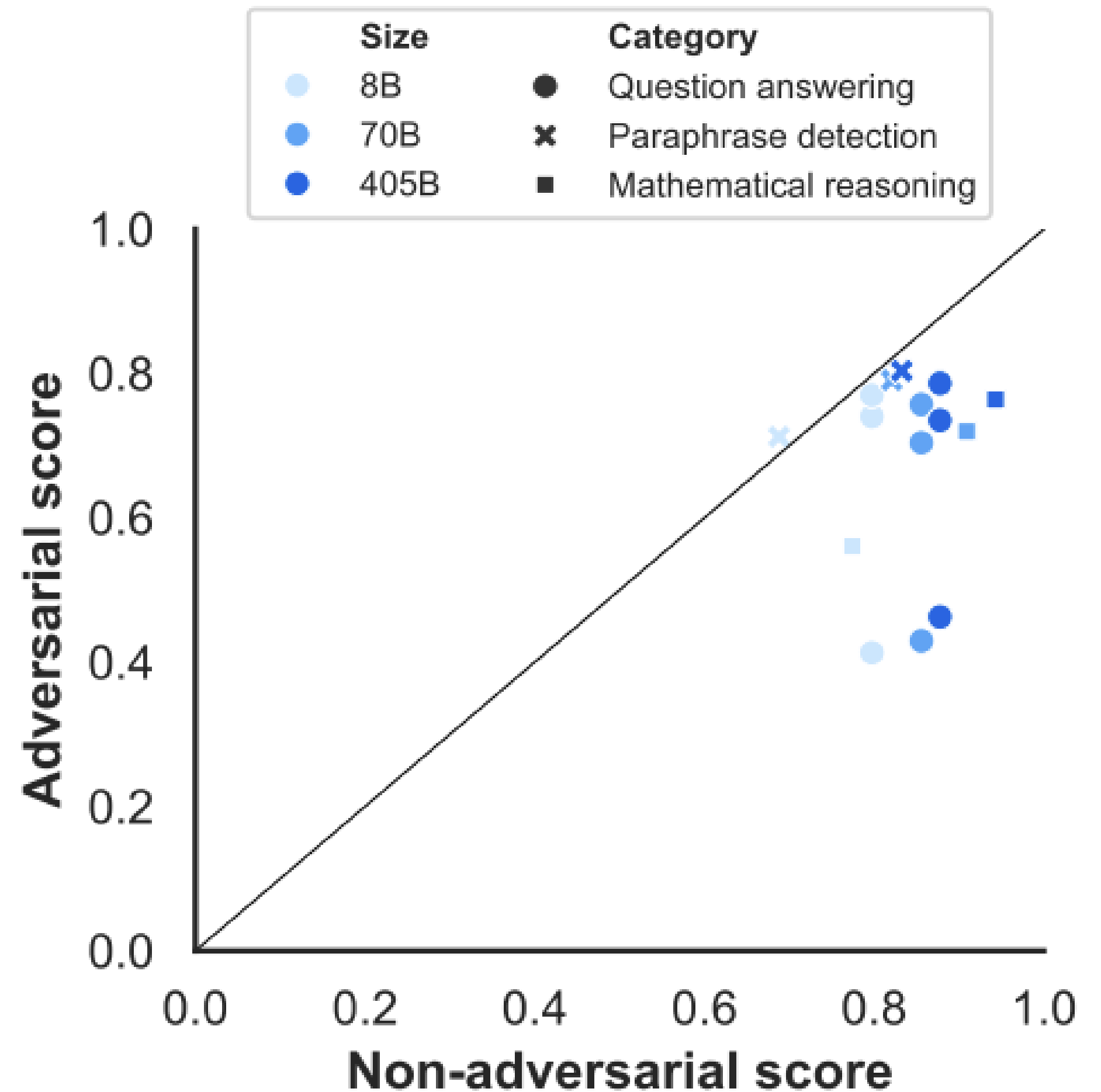
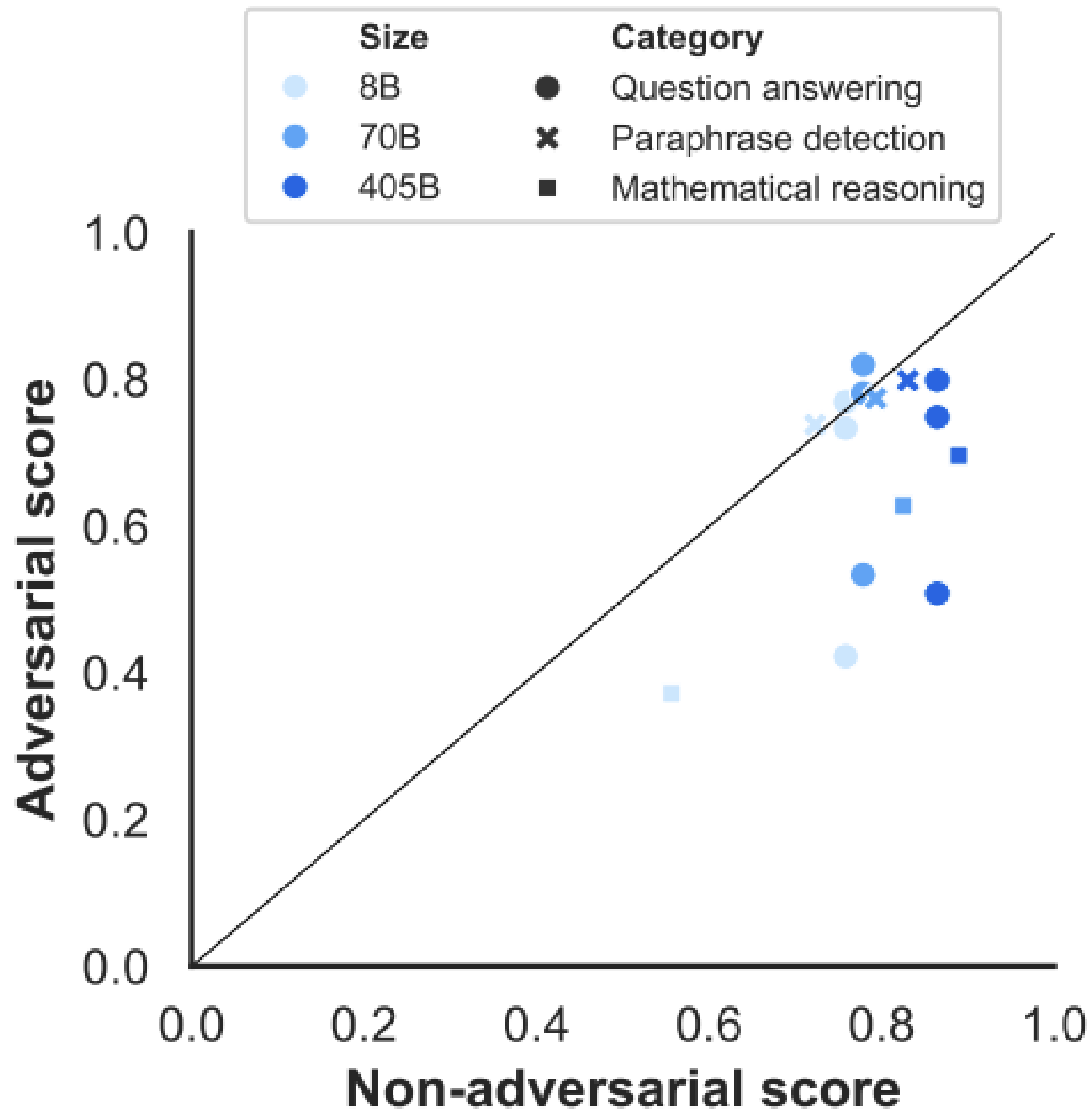
Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. [Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.](#)”*

Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

5. Pre-trained Models



Contamination Analysis

- Determine how much benchmark scores are influenced by data in pre-training corpus
- Estimated performance gain from contaminated over clean portions of a dataset
- An example of a dataset is contaminated if a ratio of its tokens overlap an 8-gram in the pre-training corpus
- The threshold for the contamination ratio is picked for each dataset to showcase the maximum performance gain

5. Pre-trained Models

	Contam.	Performance gain est.		
		8B	70B	405B
AGIEval	98	8.5	19.9	16.3
BIG-Bench Hard	95	26.0	36.0	41.0
BoolQ	96	4.0	4.7	3.9
CommonSenseQA	30	0.1	0.8	0.6
DROP	–	–	–	–
GSM8K	41	0.0	0.1	1.3
→ HellaSwag	85	14.8	14.8	14.3
HumanEval	–	–	–	–
MATH	1	0.0	-0.1	-0.2
MBPP	–	–	–	–
MMLU	–	–	–	–
MMLU-Pro	–	–	–	–
→ NaturalQuestions	52	1.6	0.9	0.8
OpenBookQA	21	3.0	3.3	2.6
→ PiQA	55	8.5	7.9	8.1
QuaC	99	2.4	11.0	6.4
RACE	–	–	–	–
SiQA	63	2.0	2.3	2.6
SQuAD	0	0.0	0.0	0.0
Winogrande	6	-0.1	-0.1	-0.2
WorldSense	73	-3.1	-0.4	3.9

5. Post-trained Models

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.

5. Post-trained Models

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	72.3	61.1	83.6	76.9	70.7	87.3	82.6	85.1	89.1	89.9
	MMLU (0-shot, CoT)	73.0	72.3 [△]	60.5	86.0	79.9	69.8	88.6	78.7 [◁]	85.4	88.7	88.3
	MMLU-Pro (5-shot, CoT)	48.3	–	36.9	66.4	56.3	49.2	73.3	62.7	64.8	74.0	77.0
	IFEval	80.4	73.6	57.6	87.5	72.7	69.9	88.6	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0	89.0	73.2	86.6	90.2	92.0
	MBPP EvalPlus (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0	88.6	72.8	83.6	87.8	90.5
Math	GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6	96.8	92.3 [◇]	94.2	96.1	96.4 [◇]
	MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1	73.8	41.1	64.5	76.6	71.1
Reasoning	ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7	96.9	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	46.7	33.3	30.8	51.1	–	41.4	53.6	59.4
Tool use	BFCL	76.1	–	60.4	84.8	–	85.9	88.5	86.5	88.3	80.5	90.2
	Nexus	38.5	30.0	24.7	56.7	48.5	37.2	58.7	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	95.2	–	95.2	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	83.4	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	100.0	100.0	90.8
Multilingual	MGSM (0-shot, CoT)	68.9	53.2	29.9	86.9	71.1	51.4	91.6	–	85.9	90.5	91.6

5. Post-trained Models

Exam	Llama 3 8B	Llama 3 70B	Llama 3 405B	GPT-3.5 Turbo	Nemotron 4 340B	GPT-4o	Claude 3.5 Sonnet
LSAT	53.9 ±4.9	74.2 ±4.3	81.1 ±3.8	54.3 ±4.9	73.7 ±4.3	77.4 ±4.1	80.0 ±3.9
SAT Reading	57.4 ±4.2	71.4 ±3.9	74.8 ±3.7	61.3 ±4.2	–	82.1 ±3.3	85.1 ±3.1
SAT Math	73.3 ±4.6	91.9 ±2.8	94.9 ±2.3	77.3 ±4.4	–	95.5 ±2.2	95.8 ±2.1
GMAT Quant.	56.0 ±19.5	84.0 ±14.4	96.0 ±7.7	36.0 ±18.8	76.0 ±16.7	92.0 ±10.6	92.0 ±10.6
GMAT Verbal	65.7 ±11.4	85.1 ±8.5	86.6 ±8.2	65.7 ±11.4	91.0 ±6.8	95.5 ±5.0	92.5 ±6.3
GRE Physics	48.0 ±11.3	74.7 ±9.8	80.0 ±9.1	50.7 ±11.3	–	89.3 ±7.0	90.7 ±6.6
AP Art History	75.6 ±12.6	84.4 ±10.6	86.7 ±9.9	68.9 ±13.5	71.1 ±13.2	80.0 ±11.7	77.8 ±12.1
AP Biology	91.7 ±11.1	100.0 ±0.0	100.0 ±0.0	91.7 ±11.1	95.8 ±8.0	100.0 ±0.0	100.0 ±0.0
AP Calculus	57.1 ±16.4	54.3 ±16.5	88.6 ±10.5	62.9 ±16.0	68.6 ±15.4	91.4 ±9.3	88.6 ±10.5
AP Chemistry	59.4 ±17.0	96.9 ±6.0	90.6 ±10.1	62.5 ±16.8	68.8 ±16.1	93.8 ±8.4	96.9 ±6.0
AP English Lang.	69.8 ±12.4	90.6 ±7.9	94.3 ±6.2	77.4 ±11.3	88.7 ±8.5	98.1 ±3.7	90.6 ±7.9
AP English Lit.	59.3 ±13.1	79.6 ±10.7	83.3 ±9.9	53.7 ±13.3	88.9 ±8.4	88.9 ±8.4	85.2 ±9.5
AP Env. Sci.	73.9 ±12.7	89.1 ±9.0	93.5 ±7.1	73.9 ±12.7	73.9 ±12.7	89.1 ±9.0	84.8 ±10.4
AP Macro Eco.	72.4 ±11.5	98.3 ±3.3	98.3 ±3.3	67.2 ±12.1	91.4 ±7.2	96.5 ±4.7	94.8 ±5.7
AP Micro Eco.	70.8 ±12.9	91.7 ±7.8	93.8 ±6.8	64.6 ±13.5	89.6 ±8.6	97.9 ±4.0	97.9 ±4.0
AP Physics	57.1 ±25.9	78.6 ±21.5	92.9 ±13.5	35.7 ±25.1	71.4 ±23.7	71.4 ±23.7	78.6 ±21.5
AP Psychology	94.8 ±4.4	100.0 ±0.0	100.0 ±0.0	94.8 ±4.4	100.0 ±0.0	100.0 ±0.0	100.0 ±0.0
AP Statistics	66.7 ±17.8	59.3 ±18.5	85.2 ±13.4	48.1 ±18.8	77.8 ±15.7	92.6 ±9.9	96.3 ±7.1
AP US Gov.	90.2 ±9.1	97.6 ±4.7	97.6 ±4.7	78.0 ±12.7	78.0 ±12.7	100.0 ±0.0	100.0 ±0.0
AP US History	78.0 ±12.7	97.6 ±4.7	97.6 ±4.7	85.4 ±10.8	70.7 ±13.9	95.1 ±6.6	95.1 ±6.6
AP World History	94.1 ±7.9	100.0 ±0.0	100.0 ±0.0	88.2 ±10.8	85.3 ±11.9	100.0 ±0.0	97.1 ±5.7
AP Average	74.1 ±3.4	87.9 ±2.5	93.5 ±1.9	70.2 ±3.5	81.3 ±3.0	93.0 ±2.0	92.2 ±2.1
GRE Quant.	152.0	158.0	162.0	155.0	161.0	166.0	164.0
GRE Verbal	149.0	166.0	166.0	154.0	162.0	167.0	167.0

5. Post-trained Models

Model	HumanEval	HumanEval+	MBPP	MBPP EvalPlus (base)
Llama 3 8B	72.6 ± 6.8	67.1 ± 7.2	60.8 ± 4.3	72.8 ± 4.5
Gemma 2 9B	54.3 ± 7.6	48.8 ± 7.7	59.2 ± 4.3	71.7 ± 4.5
Mistral 7B	40.2 ± 7.5	32.3 ± 7.2	42.6 ± 4.3	49.5 ± 5.0
Llama 3 70B	80.5 ± 6.1	74.4 ± 6.7	75.4 ± 3.8	86.0 ± 3.5
Mixtral 8 \times 22B	75.6 ± 6.6	68.3 ± 7.1	66.2 ± 4.1	78.6 ± 4.1
GPT-3.5 Turbo	68.0 ± 7.1	62.8 ± 7.4	71.2 ± 4.0	82.0 ± 3.9
Llama 3 405B	89.0 ± 4.8	82.3 ± 5.8	78.8 ± 3.6	88.6 ± 3.2
GPT-4	86.6 ± 5.2	77.4 ± 6.4	80.2 ± 3.5	83.6 ± 3.7
GPT-4o	90.2 ± 4.5	86.0 ± 5.3	81.4 ± 3.4	87.8 ± 3.3
Claude 3.5 Sonnet	92.0 ± 4.2	82.3 ± 5.8	76.6 ± 3.7	90.5 ± 3.0
Nemotron 4 340B	73.2 ± 6.8	64.0 ± 7.3	75.4 ± 3.8	72.8 ± 4.5

5. Post-trained Models

Evaluate our models on a range of benchmarks for zero-shot tool use

	Nexus	API-Bank	API-Bench	BFCL
Llama 3 8B	38.5 ± 4.1	82.6 ± 3.8	8.2 ± 1.3	76.1 ± 2.0
Gemma 2 9B	–	56.5 ± 4.9	11.6 ± 1.5	–
Mistral 7B	24.7 ± 3.6	55.8 ± 4.9	4.7 ± 1.0	60.4 ± 2.3
Llama 3 70B	56.7 ± 4.2	90.0 ± 3.0	29.7 ± 2.1	84.8 ± 1.7
Mixtral 8 \times 22B	48.5 ± 4.2	73.1 ± 4.4	26.0 ± 2.0	–
GPT-3.5 Turbo	37.2 ± 4.1	60.9 ± 4.8	36.3 ± 2.2	85.9 ± 1.7
Llama 3 405B	58.7 ± 4.1	92.3 ± 2.6	35.3 ± 2.2	88.5 ± 1.5
GPT-4	50.3 ± 4.2	89.0 ± 3.1	22.5 ± 1.9	88.3 ± 1.5
GPT-4o	56.1 ± 4.2	91.3 ± 2.8	41.4 ± 2.3	80.5 ± 1.9
Claude 3.5 Sonnet	45.7 ± 4.2	92.6 ± 2.6	60.0 ± 2.3	90.2 ± 1.4
Nemotron 4 340B	–	–	–	86.5 ± 1.6

5. Safety

- Pre-training
- Safety finetuning
- Red teaming
- System-level safety

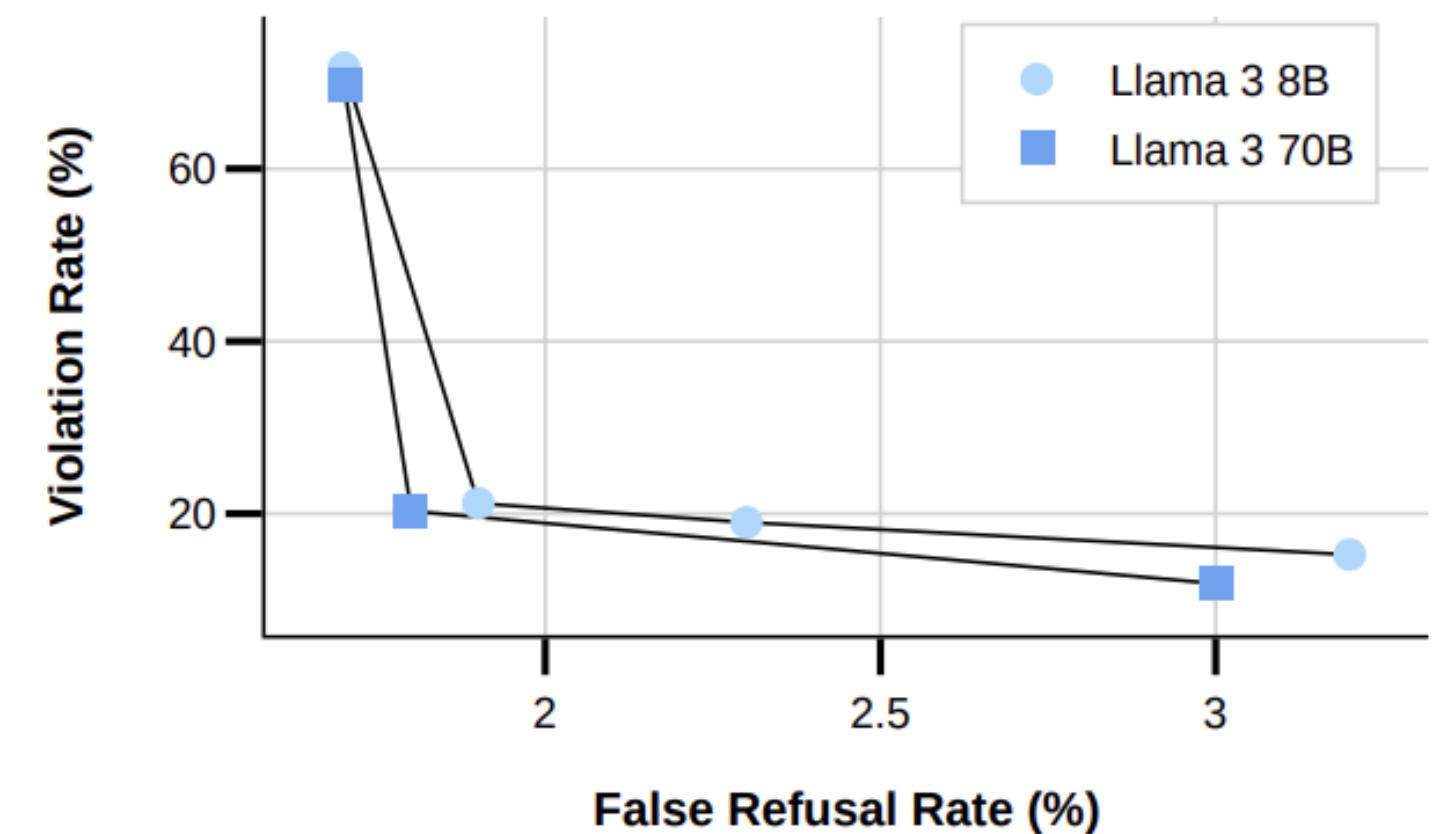
5. Safety

Benchmark Construction

- risk categories from the ML Commons taxonomy of hazards
- collect human-written prompts for each category
- 4000 per category, single- and multi-turn

Safety finetuning

- optimize for violation rate and false refusal rate (for borderline prompts)
- safety DPO



5. Safety

Uplift testing - does LLM usage provide greater threat than already existing technology like web searching

- No significant uplift for a cybersecurity challenge for experts or novices
- No significant uplift for chemical/biological weapon creation
- Similar to a study done by OpenAI, which also did not find statistically significant results

5. Safety

Red Teaming

- Adversarial testing
 - hypothetical scenarios, refusal suppression, gradually escalating
- Multilingual
 - mixing languages, language-specific slang

5. Safety

System-level safety

- Train Llama Guard 3 on 13 hazard categories
 - Training data: English data from previous iteration, multilingual, tool use

Capability	Input Llama Guard		Output Llama Guard		Full Llama Guard	
	VR	FRR	VR	FRR	VR	FRR
English	-76%	+95%	-75%	+25%	-86%	+102%
French	-38%	+27%	-45%	+4%	-59%	+29%
German	-57%	+32%	-60%	+14%	-77%	+37%
Hindi	-54%	+60%	-54%	+14%	-71%	+62%
Italian	-34%	+27%	-34%	+5%	-48%	+29%
Portuguese	-51%	+35%	-57%	+13%	-65%	+39%
Spanish	-41%	+26%	-50%	+10%	-60%	+27%
Thai	-43%	+37%	-39%	+8%	-51%	+39%

6. Inference

To make the inference with Llama 3 405 B model more efficient, two methods are used:

- **Pipeline Parallelism**

- Parallelize the model inference using BF16 precision across 16 GPUs on two machine
- Evaluate the effect of using two micro-batches in inference both during the K-V cache pre-filling stage of inference and decoding stage (4096 input tokens and 256 output tokens).

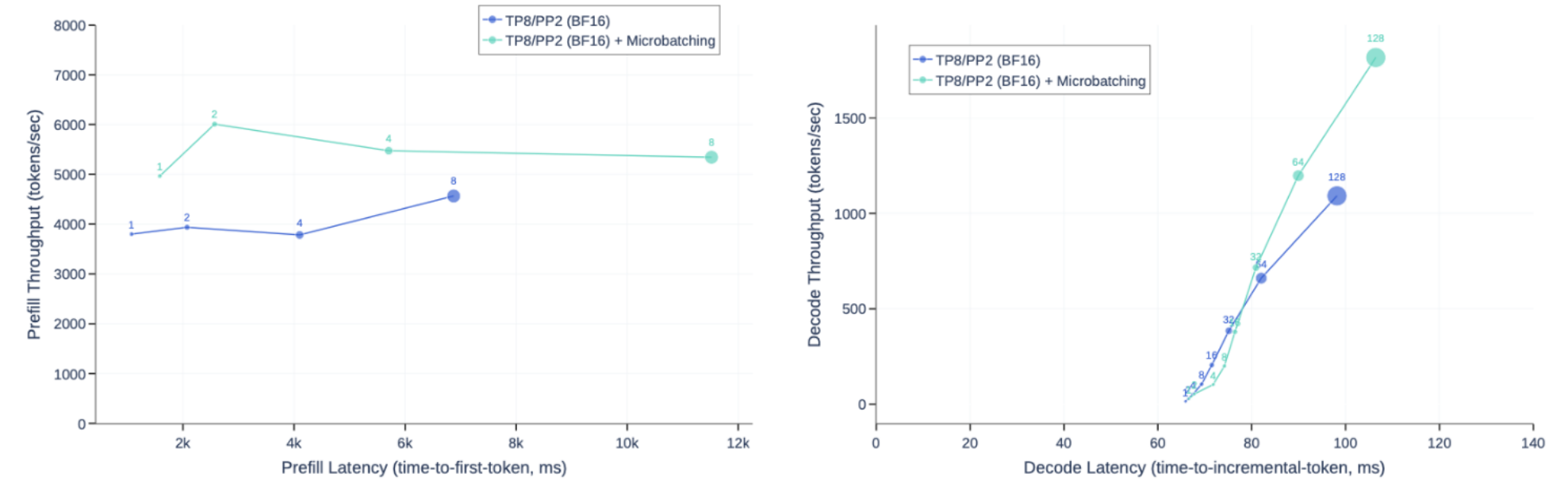


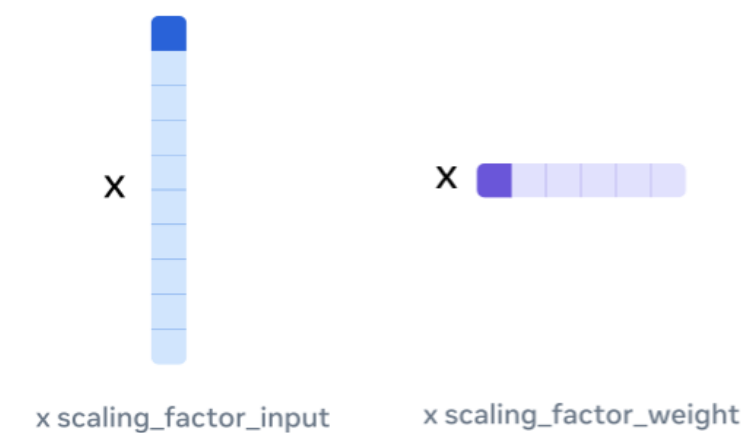
Figure 24 Effect of micro-batching on inference throughput and latency during the *Left*: pre-filling and *Right*: decoding stage. The numbers in the plot correspond to the (micro-)batch size.

micro-batching improve the inference throughput with same local batch size.

6. Inference

• FP8 Quantization

- Apply FP8 quantization in most parameters and activations in the feedforward network layers, which account for almost 50% inference time.
- Using dynamic scaling factors to improve the accuracy and set the upper bound to 1200 to prevent the error caused by high scaling factor in decoding
- Don't apply quantization in first and last layer in Transformer.
- Use row-wise quantization



, *Right:* Row-wise quantization enables the use of more

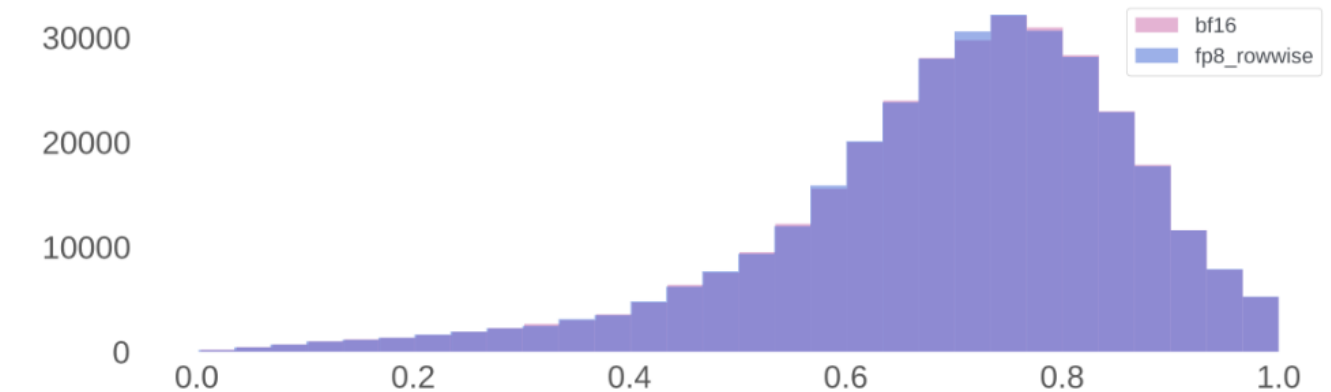


Figure 26 Reward score distribution for Llama 3 405B using BF16 and FP8 inference. Our FP8 quantization approach has negligible impact on the model's responses.

6. Inference

● FP8 Efficiency Evaluation

- throughput- latency trade-off of using FP8 in pre-filling stage and decoding stage with using 4096 input tokens and 256 output tokens.
- 50% improvement of throughput during pre-filling and better throughput during decoding

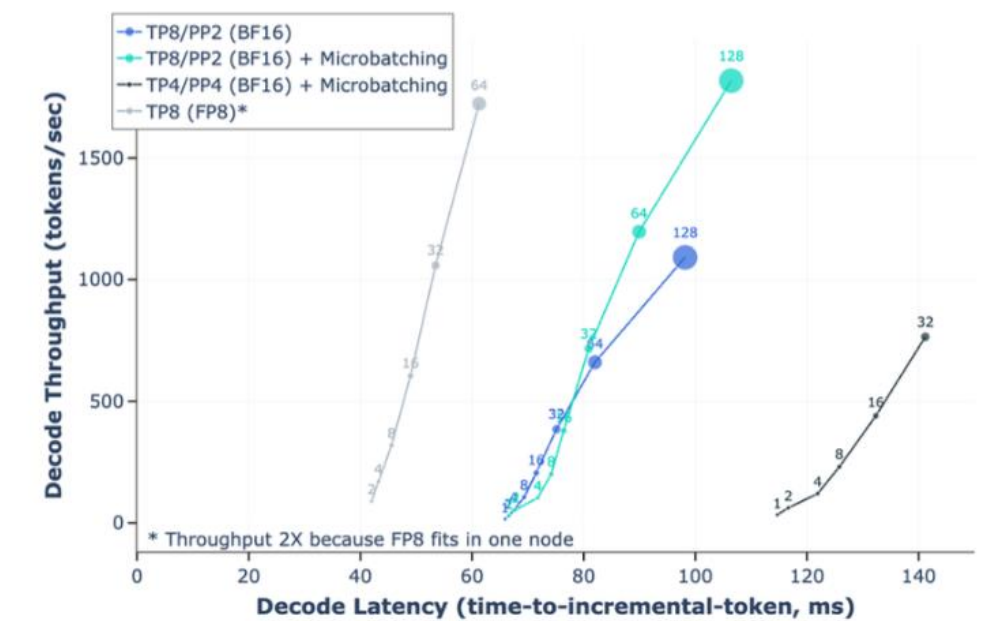
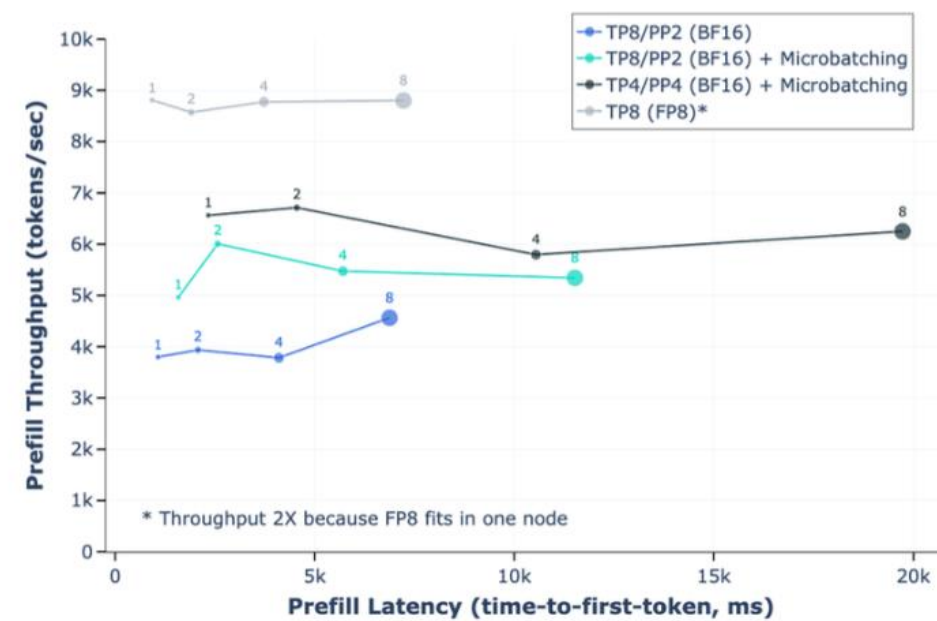


Figure 27 Throughput-latency trade-off in FP8 inference with Llama 3 405B compared with BF16 inference using different pipeline parallelization setups. *Left:* Results for pre-filling. *Right:* Results for decoding.

7. Vision Experiments

Data Preparation

Image Data

- **Image-text pairs with four steps preprocessing:**
 - **Quality filtering:** remove non-English and low-quality data below certain CLIP score
 - **De-duplication:** Compute 512-dimensional representation of images using SSIM model and perform nearest neighbor search using those embeddings and using connected-components algorithms to maintain on image-text per connected component
 - **Resampling:** Construct a vocabulary of the n-grams of high quality data and compute the frequency of each vocabulary n-gram in the dataset, if the frequency of the n-gram in caption is less than T , then we keep it. Otherwise, independent sampling each of n-grams in the caption with probability of $\sqrt{T/f_i}$, where f_i is the frequency of n-gram
 - **Optical Character recognition:** Extracting the text written in the image and concatenate it with the caption
- **Transcribing documents:** render pages from documents as images and paired images with their respective text
- **Safety:** media-risk retrieval method to identify and remove the image-text pairs that to be NSFW and blurring the face in the image
- **Annealing Data:** Resampling the image-caption pair based on the n-grams to smaller datasets and augment the dataset using additional source: visual grounding, screenshot parsing, question-answer pairs, synthetic captions

7. Vision Experiments

Data Preparation

Video Data

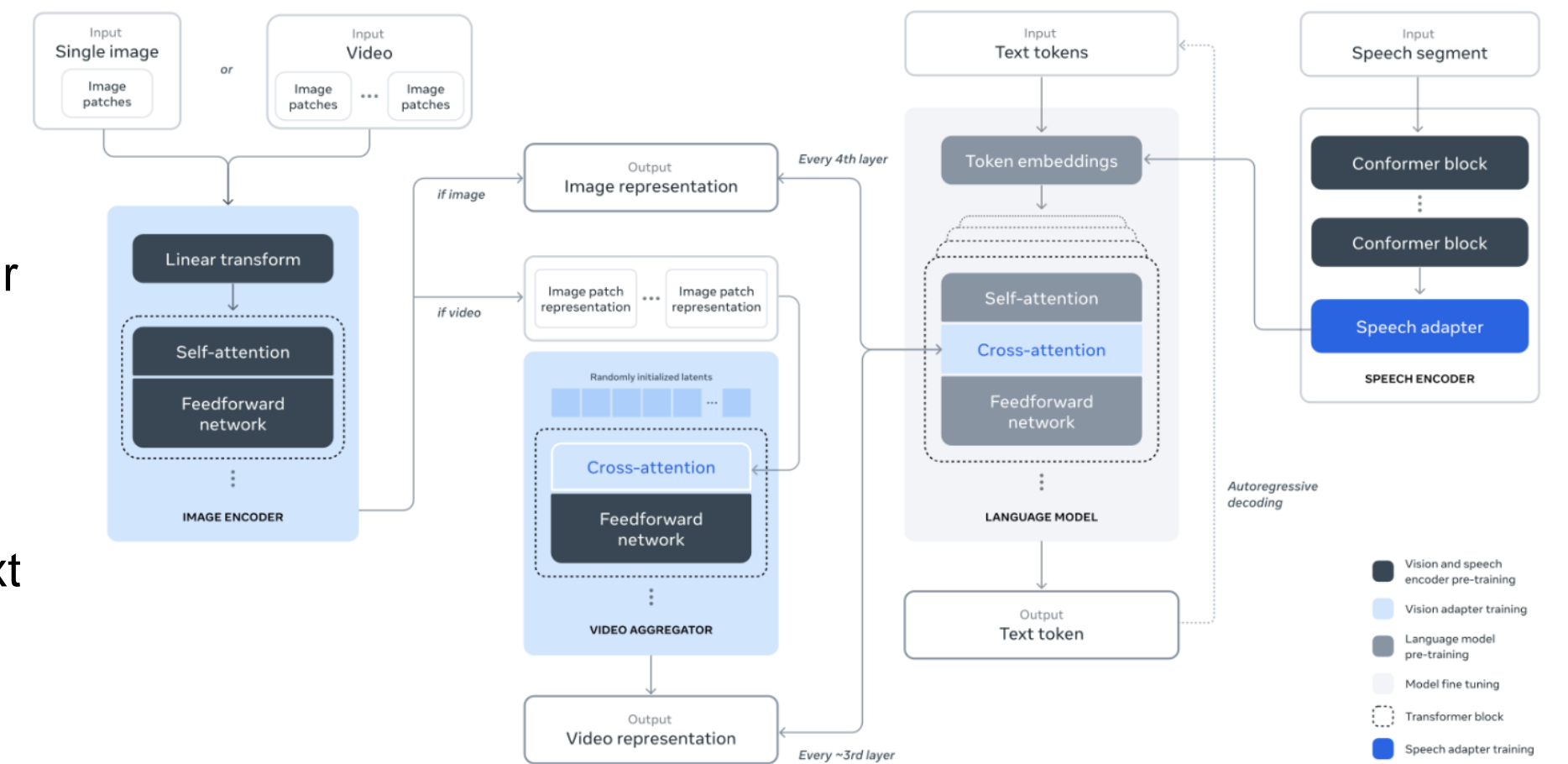
- Contain videos with an average duration of 21 seconds and a median duration of 16 seconds and most video duration is under minutes and spatial resolution varies between 320p and 4K
 - Filter and clean the associated texts to ensure a minimum length and fixing capitalization
 - Use language identification models to filter non-English texts
 - OCR detection modes to filter out video with excessive overlaid text
 - Use CLIP to ensure the video-text alignment
 - Filter out data with static or low motion using motion-score based filtering
 - Don't apply another any filter on visual quality of video

7. Vision Experiments

Model Structure

Three Main components:

- **Image encoder:**
 - ViT-H/14 variant of image encoder.
 - Images are split in to 16*16 patches
 - use multi-layer feature extraction from 4th, 8th, 16th, 24th and 31st layer to the final layer
 - Insert 8 gated self-attention layer prior to pre-training of the cross attention layer
- **Image Adaptor:**
 - cross-attention layer between the visual token and text token.
 - Apply after every fourth self-attention layer in the core language model
 - Pre-trained using 6B image-text data from image dataset and 500 M image-text data form annealing dataset for annealing model
- **Video Adaptor:** Split the video to frames uniformly and each frame are processed by image encoder
 - 32 consecutive encoded frames are merge into on using temporal aggregator
 - Add additional video cross-attention layers before every fourth images cross attention layer



7. Vision Experiments

Model Scaling

- To train smaller model which has 8B and 70 B parameters, it is efficient to use a combination of data and tensor parallelization and use pipeline parallelism to train the model
- Three challenges to train at this scale:
 - **Model Heterogeneity:** More computation is performed on some images tokens other than text tokens, causing the probability of scheduling the pipeline parallelism. And address this issue by ensuring each pipeline stages have five layers: four self-attention layers in language backbone and a cross-attention layer and replicate the image encoder on all pipeline stages
 - **Data Heterogeneity:** On average, images have more tokens than the associated text, therefore the cross-attention layer need more computational time compared with self-attention layers. Address this issue by introducing sequence parallelization in image encoder so that each GPU can process same amount of tokens
 - **Numerical Instabilities:** Gradients accumulation in bf16 lead to numerical instabilities. Therefore perform gradient accumulation in FP32.

7. Vision Experiments

Pre-training

- **Image:**

- Initialize the weights of Language Model and vision encoder, and vision encoder's weight keep unfrozen during the training.
- First train the model using 6B image-text pairs and images are resize to 336*336 pixels
- Global batch with size 16834 and initialize learning rate $10 \cdot 10^{-4}$ with weight of decay 0.01.
- After the base pre-training increase the image resolution further and train the model with the same weights for annealing dataset with re-initialized optimizer learning rate = $2 \cdot 10^{-5}$

- **Video:**

- Using the same strategy from the based pretrain and annealed image encoder
- Add and initialize randomly the video aggregator and video cross-attention layer with frozen all other weights and pre-train them in video-text pair data
- Using the same training parameters similar to image training

7. Vision Experiments

Post-training: To boost the performance of human preference evaluation

- **Supervised Fine-tuning:** Involves further training the pre-trained model on a curated set of human-annotated data or synthetic data (images, videos) to improve performance in specific tasks like multimodal conversation, image recognition, and language understanding
- **Reward Modeling:** Trains a reward model using human-annotated preference data to rank outputs (edited > chosen > rejected). This helps the model learn to prioritize higher-quality responses, improving the alignment with human preferences
- **Direct Preference Optimization:** Further train the vision-adapters with DPO using the preference data.
- **Rejection sampling:** Use the rejection sampling to generate the missing explanations for examples that lack of chain-of-thought explanations and boosts the model's reasoning ability.
- **Quality Tuning:** curate a small dataset SFT where all the samples have be rewritten and verified. And train the model after DPO process with this small dataset to improves human response quality

7. Vision Experiments

Results:

- Image:

	Llama 3-V 8B	Llama 3-V 70B	Llama 3-V 405B	GPT-4V	GPT-4o	Gemini 1.5 Pro	Claude 3.5
MMMU (val, CoT)	49.6	60.6	64.5	56.4	69.1	62.2	68.3
VQAv2 (test-dev)	78.0	79.1	80.2	77.2	–	80.2	–
AI2 Diagram (test)	84.4	93.0	94.1	78.2	94.2	94.4	94.7
ChartQA (test, CoT)	78.7	83.2	85.8	78.4	85.7	87.2	90.8
TextVQA (val)	78.2	83.4	84.8	78.0	–	78.7	–
DocVQA (test)	84.4	92.2	92.6	88.4	92.8	93.1 [△]	95.2

Table 29 Image understanding performance of our vision module attached to Llama 3. We compare model performance to GPT-4V, GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. [△]Results obtained using external OCR tools.

Outperform GPT-4 on VQAv2 and TextVQA

7. Vision Experiments

Results:

- Video

	Llama 3-V 8B	Llama 3-V 70B	Gemini 1.0 Pro	Gemini 1.0 Ultra	Gemini 1.5 Pro	GPT-4V	GPT-4o
PerceptionTest (test)	53.8	60.8	51.1	54.7	–	–	–
TVQA (val)	82.5	87.9	–	–	–	87.3	–
NExT-QA (test)	27.3	30.3	28.0	29.9	–	–	–
ActivityNet-QA (test)	52.7	56.3	49.8	52.2	57.5	–	61.9

Table 30 Video understanding performance of our vision module attached to Llama 3. We find that across range of tasks covering long-form and temporal video understanding, our vision adapters for Llama3 8B and 70B parameters are competitive and sometimes even outperform alternative models.

8. Speech Experiments

Data - Speech Understanding

- **Pre-training Data:** Curate a 15M hours of speech recordings for various Language
- **Speech recognition and translation Data:** 230K hours of manually transcribed speed recordings for 34 languages and 90K hours translation: from English to 34 languages and from 34 languages to English
- **Spoken Dialogue Data:** Synthesize 25K hours of responses for speech prompts by asking language model to answer transcripts of speech prompts

8. Speech Experiments

Data - Speech Generation

- **Text Normalization Data:** 55K pairs of written-form and corresponding speech form text for wide range of semiotic class

- **Prosody Modeling Data:** 50K hours TTS data paired by transcripts and audio recorded by professionals

8. Speech Experiments

Model Structure

- **Speech Understanding:**
 - **Speech Encoder:** Conformer with 1B parameters. The input to the model consist of 80-dimensional mel-spectrogram features and be processed by a stride-4 stacking layer and a linear projection to reduce the frame length to 40ms. And the results will be process by an encoder with 24 Conformer layers.
 - **Speech adapter:** Contains about 100 M parameters and composed of a convolutional layer, a rotary Transformer layer and a linear layer to map and match the output dimension of the language model embedding layer

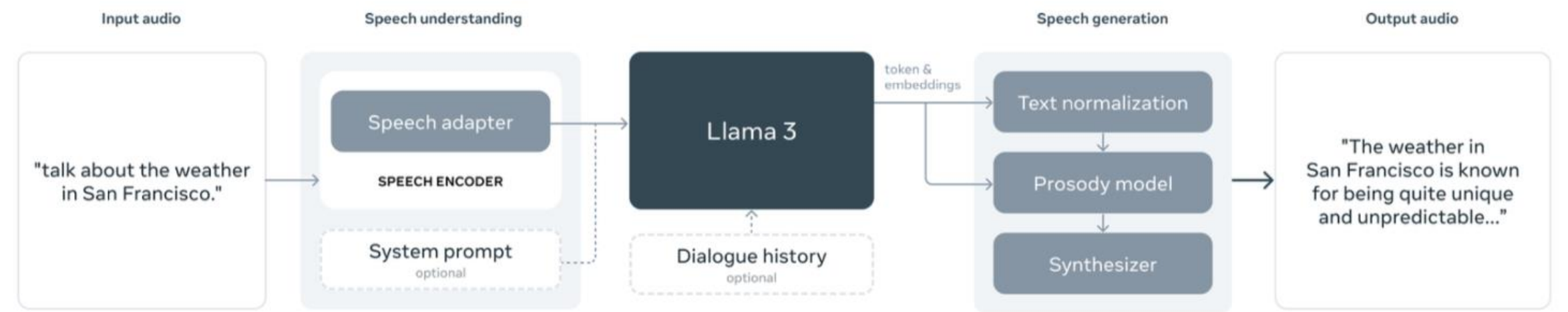


Figure 29 Architecture of our speech interface for Llama 3.

8. Speech Experiments

Model Structure

- **Speech Generation**
 - Text normalization: context-aware transformation from written-from text into corresponding spoken form using LSTM-based sequence-tagging model
 - Prosody modeling: Decoder-only Transformer based Prosody model to enhance the naturalness and expressiveness of speech synthesized



Figure 29 Architecture of our speech interface for Llama 3.

8. Speech Experiments

Training

- **Speech understanding**

- Speech pre-training: Pre-train the speech encoder using BEST-RQ algorithm
- Supervised fine-tuning: Decoder-only Transformer based Prosody model to enhance the naturalness and expressiveness of speech synthesized

- **Speech Generation:**

- Training: Using lookahead mechanism casual masking to facilitate steamability in speech synthesis
- Inference: Same mechanism and masking method are used to ensure the consistency between training and inference



Figure 29 Architecture of our speech interface for Llama 3.

8. Speech Experiments

Results:

- Speech understanding- Speech Recognition

	Llama 3 8B	Llama 3 70B	Whisper	SeamlessM4T v2	Gemini 1.0 Ultra	Gemini 1.5 Pro
MLS (English)	4.9	4.4	6.2 (v2)	6.5	4.4	4.2
LibriSpeech (test-other)	3.4	3.1	4.9 (v2)	6.2	–	–
VoxPopuli (English)	6.2	5.7	7.0 (v2)	7.0	–	–
FLEURS (34 languages)	9.6	8.2	14.4 (v3)	11.7	–	–

Table 31 Word error rate of our speech interface for Llama 3 on speech recognition tasks. We report the performance of Whisper, SeamlessM4T, and Gemini for reference.

8. Speech Experiments

Results:

- Speech understanding-Speech Translation

	Llama 3 8B	Llama 3 70B	Whisper v2	SeamlessM4T v2
FLEURS (33 lang. → English)	29.5	33.7	21.9	28.6
Covost 2 (15 lang. → English)	34.4	38.8	33.8	37.9

Table 32 BLEU score of our speech interface for Llama 3 on speech translation tasks. We report the performance of Whisper and SeamlessM4T for reference.

8. Speech Experiments

Results:

- **Speech understanding-Spoken question answering**

The figure displays three panels of transcribed dialogue examples, each labeled "audio conversation".

Panel 1 (English/German):

- User: Hey, do you know the movie "Ein Fisch namens Wanda"? Sorry, I don't know the English title
- Assistant: I think I can help you out! I believe the movie you're thinking of is called "A Fish Called Wanda" in English. It's a 1988 British comedy film directed by Charles Crichton, written by John Cleese and Charles [...]
- User: So, at some point, John Cleese says "Le due cupole grandi della cattedrale di Milano." Was heisst'n das eigentlich?
- Assistant: "Le due cupole" ist italienisch und bedeutet "die zwei Kuppeln". Die Kathedrale von Mailand, auch Duomo von Mailand genannt, ist ein berühmtes Bauwerk in Mailand, Italien, und es hat tatsächlich zwei Kuppeln: die Kuppel des Hauptbaus und die Kuppel des Campanile (Glockenturms) [...]

Panel 2 (Vietnamese/English):

- User: So I'm going to Vietnam for vacation. Ở Việt Nam chỗ nào uống cà phê ngon nhất?
- Assistant: Việt Nam là một quốc gia có nền văn hóa cà phê phong phú và đa dạng. Dưới đây là một số địa điểm uống cà phê ngon nhất ở Việt Nam: Cà phê Gò Vấp (Hồ Chí Minh): Đây là một trong những quán cà phê nổi tiếng nhất ở Việt Nam [...]
- User: Do you know what "bún chả" is?
- Assistant: I think I do! Bún chả is a popular Vietnamese dish, especially in Hanoi. It's a type of noodle dish made from rice noodles (bún) and grilled pork (chả). The dish is typically served with a sweet and savory sauce, herbs, and other toppings [...]

Panel 3 (Chinese/English):

- User: 从旧金山飞到长春要多久呢
- Assistant: 从旧金山到长春的飞行时间约为约20小时。
- User: 能骑自行车么
- Assistant: 不可以。从旧金山到长春的距离太远，需要跨越太平洋和亚太地区的多个国家。骑自行车是不可能的，因为没有陆路连接旧金山和长春。自行车也无法在海上行驶。需要通过飞机或其他交通工具才能完成这段旅程。
- User: 哈哈，开个玩笑。Do you know where can I find the best 尖椒干豆腐 after arrival?
- Assistant: 哈哈，长春的尖椒干豆腐是很有名的！如果你想找最好吃的尖椒干豆腐，以下是一些建议： [...]

Figure 30 Transcribed dialogue examples using the speech interface for Llama 3. The examples illustrate zero-shot multi-turn and code-switching capabilities.

8. Speech Experiments

Results:

- **Speech understanding-Safety**

Language	Llama 3 8B		Llama 3 70B		Gemini 1.5 Pro	
	AT (↓)	LT (↑)	AT (↓)	LT (↑)	AT (↓)	LT (↑)
English	0.84	15.09	0.68	15.46	1.44	13.42
Overall	2.31	9.89	2.00	10.29	2.06	10.94

Table 33 Speech toxicity of our speech interface to Llama 3 on the MuTox dataset. AT refers to added toxicity (%) and LT refers to lost toxicity (%).

8. Speech Experiments

Results:

- **Speech Generation - Text Normalization**

Model	Context	Accuracy
Without Llama 3 8B	3	73.6%
Without Llama 3 8B	∞	88.0%
With Llama 3 8B	3	90.7%

Table 34 Sample-wise text normalization (TN) accuracy.

We compare models with or without Llama 3 8B embeddings, and using different right-context values.

8. Speech Experiments

Results:

- **Speech Generation - Prosody Modeling**

Model	Preference
PM for Llama 3 8B	60.0%
Streaming phone-only baseline	40.0%

Model	Preference
PM for Llama 3 8B	63.6%
Non-streaming phone-only baseline	36.4%

Table 35 Prosody Modeling (PM) evaluation. *Left:* Rater preferences of PM for Llama 3 8B vs. streaming phone-only baseline. *Right:* Rater preferences of PM for Llama 3 8B vs. non-streaming phone-only baseline.



Thanks for Listening!

Georgia Tech, Atlanta