

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

**Table 3: Overall performance on grouped academic benchmarks compared to open-source base models.**

- **Popular Aggregated Benchmarks.** We report the overall results for MMLU (5 shot) (Hendrycks et al., 2020), Big Bench Hard (BBH) (3 shot) (Suzgun et al., 2022), and AGI Eval (3–5 shot) (Zhong et al., 2023). For AGI Eval, we only evaluate on the English tasks and report the average.

As shown in Table 3, LLAMA 2 models outperform LLAMA 1 models. In particular, LLAMA 2 70B improves the results on MMLU and BBH by  $\approx 5$  and  $\approx 8$  points, respectively, compared to LLAMA 1 65B. LLAMA 2 7B and 30B models outperform MPT models of the corresponding size on all categories besides code benchmarks. For the Falcon models, LLAMA 2 7B and 34B outperform Falcon 7B and 40B models on all categories of benchmarks. Additionally, LLAMA 2 70B model outperforms all open-source models.

In addition to open-source models, we also compare LLAMA 2 70B results to closed-source models. As shown in Table 4, LLAMA 2 70B is close to GPT-3.5 (OpenAI, 2023) on MMLU and GSM8K, but there is a significant gap on coding benchmarks. LLAMA 2 70B results are on par or better than PaLM (540B) (Chowdhery et al., 2022) on almost all benchmarks. There is still a large gap in performance between LLAMA 2 70B and GPT-4 and PaLM-2-L.

We also analysed the potential data contamination and share the details in Section A.6.

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	<b>86.4</b>	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	<b>86.1</b>	85.0
Natural Questions (1-shot)	–	–	29.3	<b>37.5</b>	33.0
GSM8K (8-shot)	57.1	<b>92.0</b>	56.5	80.7	56.8
HumanEval (0-shot)	48.1	<b>67.0</b>	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	<b>65.7</b>	51.2

**Table 4: Comparison to closed-source models** on academic benchmarks. Results for GPT-3.5 and GPT-4 are from OpenAI (2023). Results for the PaLM model are from Chowdhery et al. (2022). Results for the PaLM-2-L are from Anil et al. (2023).

### 3 Fine-tuning

LLAMA 2-CHAT is the result of several months of research and iterative applications of alignment techniques, including both instruction tuning and RLHF, requiring significant computational and annotation resources.

In this section, we report on our experiments and findings using supervised fine-tuning (Section 3.1), as well as initial and iterative reward modeling (Section 3.2.2) and RLHF (Section 3.2.3). We also share a new technique, Ghost Attention (GAt), which we find helps control dialogue flow over multiple turns (Section 3.3). See Section 4.2 for safety evaluations on fine-tuned models.