
Mixtral of Experts

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch,
Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,
Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour,
Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux,
Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao,
Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, William El Sayed



Abstract

We introduce Mixtral 8x7B, a Sparse Mixture of Experts (SMoE) language model. Mixtral has the same architecture as Mistral 7B, with the difference that each layer is composed of 8 feedforward blocks (i.e. experts). For every token, at each layer, a router network selects two experts to process the current state and combine their outputs. Even though each token only sees two experts, the selected experts can be different at each timestep. As a result, each token has access to 47B parameters, but only uses 13B active parameters during inference. Mixtral was trained with a context size of 32k tokens and it outperforms or matches Llama 2 70B and GPT-3.5 across all evaluated benchmarks. In particular, Mixtral vastly outperforms Llama 2 70B on mathematics, code generation, and multilingual benchmarks. We also provide a model fine-tuned to follow instructions, Mixtral 8x7B – Instruct, that surpasses GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B – chat model on human benchmarks. Both the base and instruct models are released under the Apache 2.0 license.

Code: <https://github.com/mistralai/mistral-src>

Webpage: <https://mistral.ai/news/mixtral-of-experts/>

1 Introduction

In this paper, we present Mixtral 8x7B, a sparse mixture of experts model (SMoE) with open weights, licensed under Apache 2.0. Mixtral outperforms Llama 2 70B and GPT-3.5 on most benchmarks. As it only uses a subset of its parameters for every token, Mixtral allows faster inference speed at low batch-sizes, and higher throughput at large batch-sizes.

Mixtral is a sparse mixture-of-experts network. It is a decoder-only model where the feedforward block picks from a set of 8 distinct groups of parameters. At every layer, for every token, a router network chooses two of these groups (the “experts”) to process the token and combine their output additively. This technique increases the number of parameters of a model while controlling cost and latency, as the model only uses a fraction of the total set of parameters per token.

Mixtral is pretrained with multilingual data using a context size of 32k tokens. It either matches or exceeds the performance of Llama 2 70B and GPT-3.5, over several benchmarks. In particular,

专家混搭

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Thimoth e Lacroix, William El Sayed



摘要

我们介绍了Mixtral 8x7B，这是一种稀疏专家混合模型（SMoE）的语言模型。Mixtral与Mistral 7B具有相同的架构，不同之处在于每个层由8个前馈块组成（即专家）。对于每一个令牌，在每一层中，一个路由器网络选择两个专家来处理当前状态，并将它们的输出进行组合。尽管每个令牌只看到两个专家，但在每一步的时间步中，所选的专家可能会有所不同。因此，每个令牌可以访问47亿个参数，但仅在推理过程中使用13亿个活跃参数。Mixtral通过训练集大小为32k令牌的方式进行了训练，并且在所有评估基准上超过了或匹配了Llama 2 70B和GPT-3.5。特别是，在数学、代码生成和多语言基准中，Mixtral远超Llama 2 70B。我们还提供了一个模型微调版本，名为Mixtral 8x7B - Instruct，该模型在遵循指令的人类基准上超越了GPT-3.5 Turbo、Claude-2.1和Gemini Pro以及Llama 2 70B - chat模型。两个基础模型和微调模型都以Apache 2.0许可发布。

Code: <https://github.com/mistralai/mistral-src> Webpage: <https://mistral.ai/news/mixtral-of-experts/>

1 引言

在本论文中，我们提出了一种稀疏专家混合模型（SMoE），该模型具有开放权重，并且受Apache 2.0许可证的约束。Mixtral在大多数基准测试上优于Llama 2 70B和GPT-3.5。由于它仅使用每token的一小部分参数，Mixtral可以在低批次大小时提供更快的推理速度，并且在大规模批次大小时具有更高的吞吐量。

Mixtral是一种稀疏混合专家网络。它是解码器模型，其中前向块从一组8个不同的参数组中选择。在每一层中，对于每个令牌，路由器网络会选择两个这些组（称为“专家”）来处理该令牌并将其输出相加。这种技术增加了模型的参数数量，同时控制成本和延迟，因为模型仅使用每令牌总参数集的少量部分。

Mixtral是通过使用大小为32K块的上下文数据进行预训练的。它要么匹配，要么超过LLaMA 2 70B和GPT-3.5在多个基准测试中的表现。特别是，

4
2
0
2
n
a
J
8
|
G
L
.
s
c
|
1
v
8
8
0
4
0
.
1
0
4
2
:
v
i
X
r
a

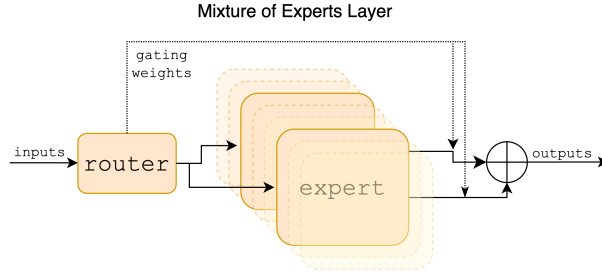


Figure 1: Mixture of Experts Layer. Each input vector is assigned to 2 of the 8 experts by a router. The layer’s output is the weighted sum of the outputs of the two selected experts. In Mixtral, an expert is a standard feedforward block as in a vanilla transformer architecture.

Mixtral demonstrates superior capabilities in mathematics, code generation, and tasks that require multilingual understanding, significantly outperforming Llama 2 70B in these domains. Experiments show that Mixtral is able to successfully retrieve information from its context window of 32k tokens, regardless of the sequence length and the location of the information in the sequence.

We also present Mixtral 8x7B – Instruct, a chat model fine-tuned to follow instructions using supervised fine-tuning and Direct Preference Optimization [25]. Its performance notably surpasses that of GPT-3.5 Turbo, Claude-2.1, Gemini Pro, and Llama 2 70B – chat model on human evaluation benchmarks. Mixtral – Instruct also demonstrates reduced biases, and a more balanced sentiment profile in benchmarks such as BBQ, and BOLD.

We release both Mixtral 8x7B and Mixtral 8x7B – Instruct under the Apache 2.0 license¹, free for academic and commercial usage, ensuring broad accessibility and potential for diverse applications. To enable the community to run Mixtral with a fully open-source stack, we submitted changes to the vLLM project, which integrates Megablocks CUDA kernels for efficient inference. Skypilot also allows the deployment of vLLM endpoints on any instance in the cloud.

2 Architectural details

Mixtral is based on a transformer architecture [31] and uses the same modifications as described in [18], with the notable exceptions that Mixtral supports a fully dense context length of 32k tokens, and the feedforward blocks are replaced by Mixture-of-Expert layers (Section 2.1). The model architecture parameters are summarized in Table 1.

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
context_len	32768
vocab_size	32000
num_experts	8
top_k_experts	2

Table 1: Model architecture.

2.1 Sparse Mixture of Experts

We present a brief overview of the Mixture of Experts layer (Figure 1). For a more in-depth overview, see [12]. The output of the MoE module for a given input x is determined by the weighted sum of the outputs of the expert networks, where the weights are given by the gating network’s output. i.e. given n expert networks $\{E_0, E_i, \dots, E_{n-1}\}$, the output of the expert layer is given by:

$$\sum_{i=0}^{n-1} G(x)_i \cdot E_i(x).$$

Here, $G(x)_i$ denotes the n -dimensional output of the gating network for the i -th expert, and $E_i(x)$ is the output of the i -th expert network. If the gating vector is sparse, we can avoid computing the outputs of experts whose gates are zero. There are multiple alternative ways of implementing $G(x)$ [6, 15, 35], but a simple and performant one is implemented by taking the softmax over the Top-K logits of a linear layer [28]. We use

$$G(x) := \text{Softmax}(\text{TopK}(x \cdot W_g)),$$

where $(\text{TopK}(\ell))_i := \ell_i$ if ℓ_i is among the top-K coordinates of logits $\ell \in \mathbb{R}^n$ and $(\text{TopK}(\ell))_i := -\infty$ otherwise. The value of K – the number of experts used per token – is a hyper-parameter that modulates the amount of compute used to process each token. If one increases n while keeping K fixed, one

¹<https://mistral.ai/news/mixtral-of-experts/>

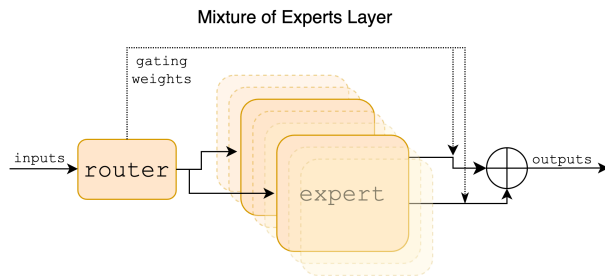


图1: 专家混合层。每个输入向量由路由器分配到8个专家中的2个。该层的输出是选择的两个专家输出的加权和。在Mixtral中, 专家是一个类似于标准前馈块的自回归架构中的普通变压器架构。

Mixtral 在数学、代码生成以及需要多语言理解的任务中展示了卓越的能力, 显著优于 Llama 2 70B 在这些领域。实验表明, Mixtral 能够成功从其 32k 个令牌的上下文中检索信息, 无论序列长度如何, 信息在序列中的位置如何。

我们还展示了 Mixtral 8x7B – Instruct, 这是一种通过监督微调和直接偏好优化 ([25]) 进行微调的聊天模型。其性能显著超过了 GPT-3.5 Turbo、Claude-2.1、Gemini Pro 和 Llama 2 70B——在人类评估基准上的聊天模型。Mixtral – Instruct 还展示了减少的偏见, 并且在基准测试中, 如 BBQ 和 BOLD 中具有更平衡的情感倾向。

我们同时发布了 Mixtral 8x7B 和 Mixtral 8x7B – Instruct, 均在 Apache 2.0 许可下¹下提供, 免费供学术和商业用途使用, 确保了广泛的可访问性和多样化的应用潜力。为了使社区能够以完全开源的堆栈运行 Mixtral, 我们向 vLLM 项目提交了更改, 该项目整合了 Megablocks CUDA 核函数以实现高效推理。Skypilot 还允许在云上的任何实例上部署 vLLM 端点。

2 建筑细节

Mixtral 是基于变压器架构 [31] 开发的, 并且使用了与文献 [18] 中描述相同的变化, 但值得注意的是 Mixtral 支持一个完全密集的上下文长度为 32k 的令牌数, 并且前向层块被混合专家层 (第 2.1 节) 所取代。模型架构参数总结见表 1。

2.1 稀疏专家混合体

我们对混合专家层 (图1) 进行了简要概述。对于更深入的概述, 请参见 [12]。给定输入 x , MoE 模块的输出由专家网络的输出按权重之和确定, 其中权重由门控网络的输出给出。即, 给定 n 专家网络 $\{E_0, E_i, \dots, E_{n-1}\}$, 专家层的输出为:

$$\sum_{i=0}^{n-1} G(x)_i \cdot E_i(x).$$

这里, $G(x)_i$ 表示门控网络对第 i -专家的 n -维输出, 而 $E_i(x)$ 是第 i -专家网络的输出。如果门控向量稀疏, 则我们可以避免计算那些门控为零的专家的输出。 $G(x)$ 的实现方式有多种选择 [6, 15, 35], 但简单且高效的实现是通过线性层的 Top-K 特征值进行 Softmax 实现的 [28]。我们使用

$$G(x) := \text{Softmax}(\text{TopK}(x \cdot W_g)),$$

where $(\text{TopK}(\ell))_i$ if ℓ_i is among the top-K coordinates of logits $\ell \in \mathbb{R}^n$ and $(\text{TopK}(\ell))_i$ otherwise. The value of K – the number of experts used per token – is a hyper-parameter that modulates the amount of compute used to process each token. If one increases n while keeping K fixed, one

¹<https://mistral.ai/news/mixtral-of-experts/>

Parameter	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
context_len	32768
vocab_size	32000
num_experts	8
top_k_experts	2

表 1: 模型架构。

can increase the model’s parameter count while keeping its computational cost effectively constant. This motivates a distinction between the model’s total parameter count (commonly referenced as the **sparse** parameter count), which grows with n , and the number of parameters used for processing an individual token (called the **active** parameter count), which grows with K up to n .

MoE layers can be run efficiently on single GPUs with high performance specialized kernels. For example, Megablocks [13] casts the feed-forward network (FFN) operations of the MoE layer as large sparse matrix multiplications, significantly enhancing the execution speed and naturally handling cases where different experts get a variable number of tokens assigned to them. Moreover, the MoE layer can be distributed to multiple GPUs through standard Model Parallelism techniques, and through a particular kind of partitioning strategy called Expert Parallelism (EP) [28]. During the MoE layer’s execution, tokens meant to be processed by a specific expert are routed to the corresponding GPU for processing, and the expert’s output is returned to the original token location. Note that EP introduces challenges in load balancing, as it is essential to distribute the workload evenly across the GPUs to prevent overloading individual GPUs or hitting computational bottlenecks.

In a Transformer model, the MoE layer is applied independently per token and replaces the feed-forward (FFN) sub-block of the transformer block. For Mixtral we use the same SwiGLU architecture as the expert function $E_i(x)$ and set $K = 2$. This means each token is routed to two SwiGLU sub-blocks with different sets of weights. Taking this all together, the output y for an input token x is computed as:

$$y = \sum_{i=0}^{n-1} \text{Softmax}(\text{Top2}(x \cdot W_g))_i \cdot \text{SwiGLU}_i(x).$$

This formulation is similar to the GShard architecture [21], with the exceptions that we replace all FFN sub-blocks by MoE layers while GShard replaces every other block, and that GShard uses a more elaborate gating strategy for the second expert assigned to each token.

3 Results

We compare Mixtral to Llama, and re-run all benchmarks with our own evaluation pipeline for fair comparison. We measure performance on a wide variety of tasks categorized as follow:

- **Commonsense Reasoning (0-shot):** Hellaswag [32], Winogrande [26], PIQA [3], SIQA [27], OpenbookQA [22], ARC-Easy, ARC-Challenge [8], CommonsenseQA [30]
- **World Knowledge (5-shot):** NaturalQuestions [20], TriviaQA [19]
- **Reading Comprehension (0-shot):** BoolQ [7], QuAC [5]
- **Math:** GSM8K [9] (8-shot) with maj@8 and MATH [17] (4-shot) with maj@4
- **Code:** Humaneval [4] (0-shot) and MBPP [1] (3-shot)
- **Popular aggregated results:** MMLU [16] (5-shot), BBH [29] (3-shot), and AGI Eval [34] (3-5-shot, English multiple-choice questions only)

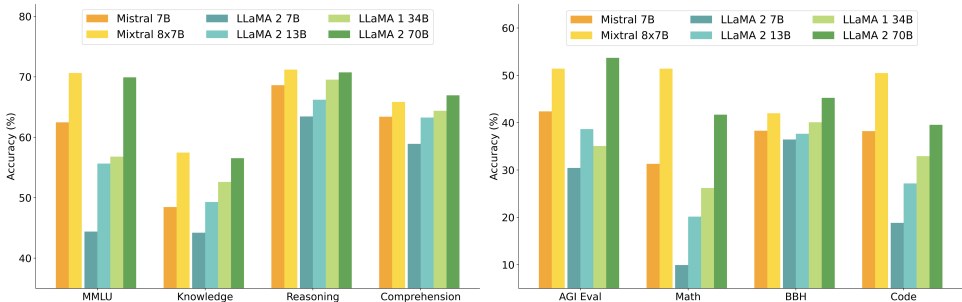


Figure 2: Performance of Mixtral and different Llama models on a wide range of benchmarks. All models were re-evaluated on all metrics with our evaluation pipeline for accurate comparison. Mixtral outperforms or matches Llama 2 70B on all benchmarks. In particular, it is vastly superior in mathematics and code generation.

可以增加模型的参数数量，同时保持其计算成本基本恒定。这促使我们区分了模型的总参数数量（通常称为稀疏参数数量），随着 n 增长；以及处理单个词元所需的参数数量（被称为活跃参数数量），随着 K 到 n 的增长。

MoE 层可以在单个 GPU 上高效运行，性能优异的专门化核。例如，Megablocks [13] 将 MoE 层中的前向传播网络（FFN）操作转换为大规模稀疏矩阵乘法，显著提高了执行速度，并自然地处理了不同专家分配给它们的令牌数量变量的情况。此外，MoE 层可以通过标准模型并行技术分布在多个 GPU 上，并通过特定类型的分区策略称为专家并行（EP）[28]。在 MoE 层执行过程中，旨在由特定专家处理的令牌被路由到相应的 GPU 进行处理，并且专家的输出返回到原始令牌位置。需要注意的是，EP 引入了负载均衡挑战，因为它必须确保均匀分配工作量以防止单个 GPU 超载或遇到计算瓶颈。

在 Transformer 模型中，MoE 层独立应用于每个令牌，并替换变压器块中的前馈（FFN）子模块。对于 Mixtral，我们使用与专家函数相同的 SwiGLU 架构 $E_i(x)$ 并设置 $K = 2$ 。这意味着每个令牌被路由到两个具有不同权重集的 SwiGLU 子模块中。将所有这些因素结合起来，输入令牌 x 的输出 y 计算如下：

$$y = \sum_{i=0}^{n-1} \text{Softmax}(\text{Top2}(x \cdot W_g))_i \cdot \text{SwiGLU}_i(x).$$

该公式与 GShard 架构类似 [21]，区别在于我们用 MoE 层替换了所有 FFN 子块，而 GShard 则将每个其他块替换为其他块，并且 GShard 对第二专家分配给每个令牌的门控策略更为复杂。

三结果

我们对 Mixtral 与 Llama，并使用我们的评估管道重新运行所有基准测试，以实现公平的比较。我们将性能测量在以下分类的任务上进行：

- 常识推理 (零样本): Hellaswag [32], Winogrande [26], PIQA [3], SIQA [27], OpenbookQA [22], ARC-Easy, ARC-Challenge [8], CommonsenseQA [30]
- 世界知识 (5次样本) : 自然问题 [20], triviaqa [19]
- 阅读理解 (零样本学习) : BoolQ [7], QuAC [5]
- 数学: GSM8K [9] (8发) 与 maj@8 和 MATH [17] (4发) 与 maj@4
- 代码: Humaneval [4] (零样本) 和 MBPP [1] (三样本)
- 热门聚合结果: MMLU [16] (5-样本), BBH [29] (3-样本), 以及 AGI 评估 [34] (3-5-样本, 仅限英语多项选择题)

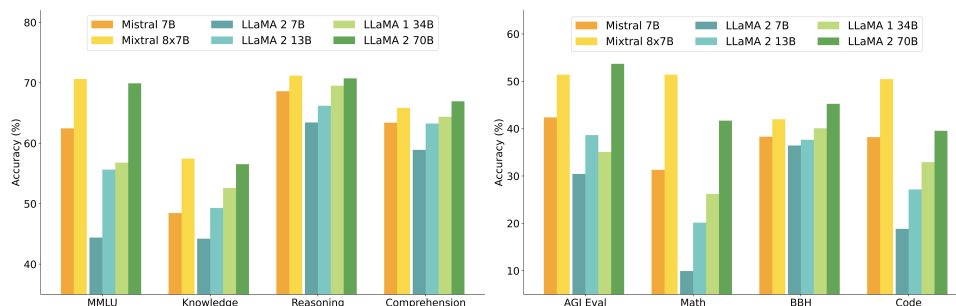


图 2: Mixtral 在广泛基准上的性能表现。所有模型均在我们的评估管道上重新评估了所有指标，以实现准确的比较。Mixtral 在所有基准测试中都优于或与 Llama 2 70B 相媲美。特别是，在数学和代码生成方面，它具有显著优势。

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	85.4%	80.4%	82.6%	79.9%	56.5%	25.4%	73.0%	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	70.6%	84.4%	77.2%	83.6%	83.1%	59.7%	30.6%	71.5%	40.2%	60.7%	28.4%	74.4%

Table 2: Comparison of Mixtral with Llama. Mixtral outperforms or matches Llama 2 70B performance on almost all popular benchmarks while using 5x fewer active parameters during inference.

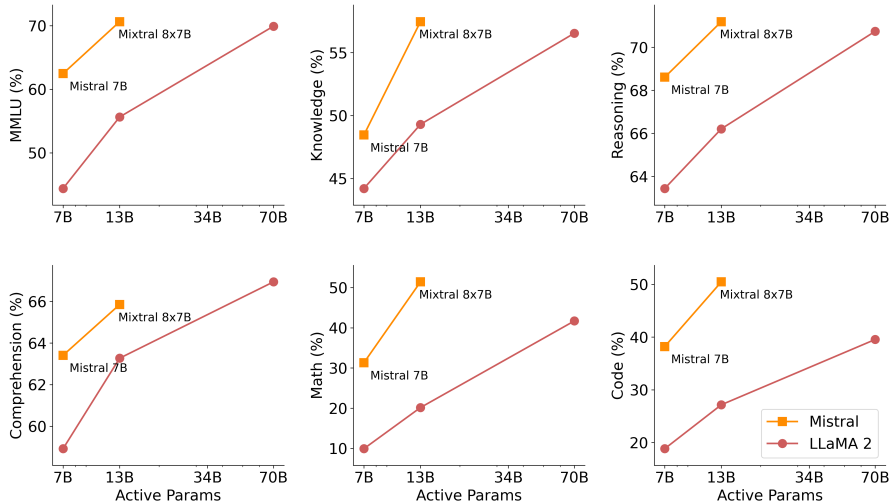


Figure 3: Results on MMLU, commonsense reasoning, world knowledge and reading comprehension, math and code for Mistral (7B/8x7B) vs Llama 2 (7B/13B/70B). Mixtral largely outperforms Llama 2 70B on all benchmarks, except on reading comprehension benchmarks while using 5x lower active parameters. It is also vastly superior to Llama 2 70B on code and math.

Detailed results for Mixtral, Mistral 7B and Llama 2 7B/13B/70B and Llama 1 34B² are reported in Table 2. Figure 2 compares the performance of Mixtral with the Llama models in different categories. Mixtral surpasses Llama 2 70B across most metrics. In particular, Mixtral displays a superior performance in code and mathematics benchmarks.

Size and Efficiency. We compare our performance to the Llama 2 family, aiming to understand Mixtral models’ efficiency in the cost-performance spectrum (see Figure 3). As a sparse Mixture-of-Experts model, Mixtral only uses 13B active parameters for each token. With 5x lower active parameters, Mixtral is able to outperform Llama 2 70B across most categories.

Note that this analysis focuses on the active parameter count (see Section 2.1), which is directly proportional to the inference compute cost, but does not consider the memory costs and hardware utilization. The memory costs for serving Mixtral are proportional to its *sparse* parameter count, 47B, which is still smaller than Llama 2 70B. As for device utilization, we note that the SMOEs layer introduces additional overhead due to the routing mechanism and due to the increased memory loads when running more than one expert per device. They are more suitable for batched workloads where one can reach a good degree of arithmetic intensity.

Comparison with Llama 2 70B and GPT-3.5. In Table 3, we report the performance of Mixtral 8x7B compared to Llama 2 70B and GPT-3.5. We observe that Mixtral performs similarly or above the two other models. On MMLU, Mixtral obtains a better performance, despite its significantly smaller capacity (47B tokens compared to 70B). For MT Bench, we report the performance of the latest GPT-3.5-Turbo model available, gpt-3.5-turbo-1106.

²Since Llama 2 34B was not open-sourced, we report results for Llama 1 34B.

Model	Active Params	MMLU	HellaS	WinoG	PIQA	Arc-e	Arc-c	NQ	TriQA	HumanE	MBPP	Math	GSM8K
LLaMA 2 7B	7B	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	17.5%	56.6%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	13B	55.6%	80.7%	72.9%	80.8%	75.2%	48.8%	16.7%	64.0%	18.9%	35.4%	6.0%	34.3%
LLaMA 1 33B	33B	56.8%	83.7%	76.2%	82.2%	79.6%	54.4%	24.1%	68.5%	25.0%	40.9%	8.4%	44.1%
LLaMA 2 70B	70B	69.9%	85.4%	80.4%	82.6%	79.9%	56.5%	25.4%	73.0%	29.3%	49.8%	13.8%	69.6%
Mistral 7B	7B	62.5%	81.0%	74.2%	82.2%	80.5%	54.9%	23.2%	62.5%	26.2%	50.2%	12.7%	50.0%
Mixtral 8x7B	13B	70.6%	84.4%	77.2%	83.6%	83.1%	59.7%	30.6%	71.5%	40.2%	60.7%	28.4%	74.4%

表 2: Mistral 与 Llama 的比较。在几乎所有流行的基准测试中, Mistral 能够超过或等于 Llama 2 70B 的性能, 同时在推理过程中使用了 5 倍于参数的少。

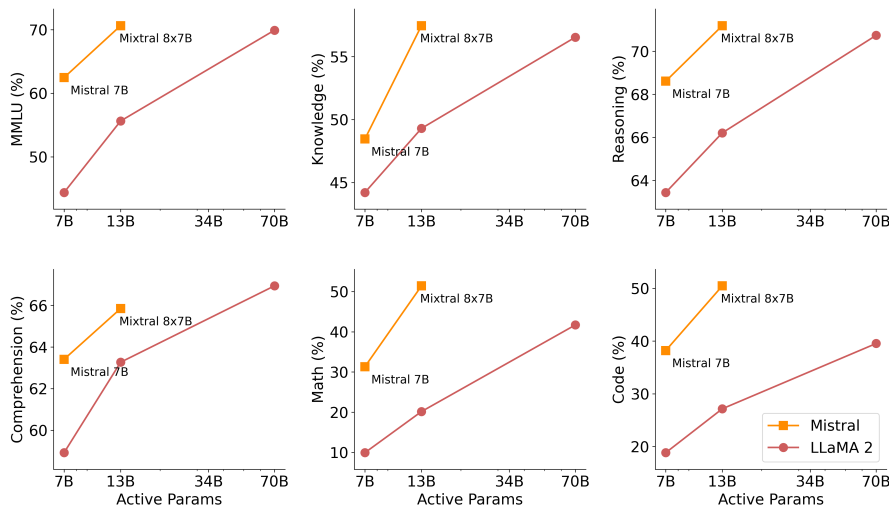


Figure 3: Mistral (7B/8x7B) 与 Llama 2 (7B/13B/70B) 在 MMLU、常识推理、世界知识和阅读理解、数学和代码方面的结果。Mistral 在所有基准测试中均优于 Llama 2 70B, 除了阅读理解基准外, 使用了 5 倍更低的活跃参数。此外, 在代码和数学方面也远胜于 Llama 2 70B。

详细结果报告了 Mistral、Mistral-7B 和 LLaMA-2 7B/13B/70B/LaMa1 34B² 的详细结果。在表 2 中, 图 2 比较了 Mistral 在不同类别中的性能与 LLaMA 模型的对比情况。Mistral 在大多数指标上超越了 LLaMA-2 70B。特别是, 在代码和数学基准测试中, Mistral 显示出更优异的表现。

大小和效率。我们与 Llama 2 家族进行比较, 旨在理解 Mistral 模型在成本-性能谱中的效率 (参见图 3)。作为稀疏专家混合模型, Mistral 每个令牌仅使用 13 亿个活跃参数。通过减少 5 倍的活跃参数, Mistral 能够在大多数类别中超越 Llama 2 70B。

注意, 本分析关注的是活跃参数计数 (见第 2.1 节), 与推理计算成本成正比, 但不考虑内存成本和硬件利用率。服务 Mistral 的内存成本与其 *sparse* 参数数量相关, 即 47B, 仍小于 Llama 2 70B。至于设备利用率, 我们注意到 SMoEs 层由于路由机制引入额外开销, 并且当在一个设备上运行多个专家时, 内存负载增加导致硬件利用率更高。它们更适合批处理工作负载, 在这种情况下可以达到良好的算术强度。

比较与 LLaMA 2 70B 和 GPT-3.5。在表 3 中, 我们报告了 Mistral 8x7B 在与 LLaMA 2 70B 和 GPT-3.5 的性能对比中的表现。观察到 Mistral 的表现与其其他两个模型相当或更好。在 MMLU 上, Mistral 获得了更好的性能, 尽管其容量显著较小 (47B 令牌对 70B)。对于 M T Bench, 我们报告了最新可用的 GPT-3.5-Turbo 模型的表现, gpt-3.5-turbo-1106。

²Since Llama 2 34B was not open-sourced, we report results for Llama 1 34B.