Phi-2

In this guide, we provide an overview of the Phi-2, a 2.7 billion parameter language model, how to prompt Phi-2, and its capabilities. This guide also includes tips, applications, limitations, important references, and additional reading materials related to Phi-2 LLM.

Phi-2 Introduction

Phi-2 is the latest small language model (SLM) released by Microsoft Research. Phi-2 follows the previous Phi-1 model and Phi-1.5 models.

Phi-1 is a 1.3 billion parameters model trained on "textbook quality" data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens) (Gunasekar et al. 2023). It performs well on Python code generation tasks.

Phi-1.5 builds on the previous model and focuses on common sense reasoning and language understanding capabilities. Phi-1.5 is capable of performing complex reasoning tasks such as grade-school mathematics and basic coding tasks, and is comparable to models 5 times larger.

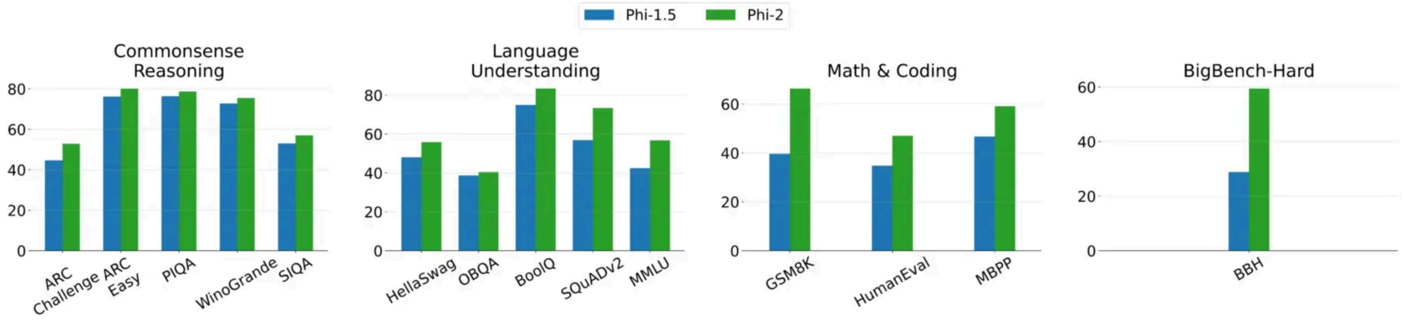
Phi-2, a 2.7 billion parameters model, improves reasoning and language understanding capabilities. Phi-2 outperforms models up to 25x larger and now has an MIT License that makes it usable in commercial settings.

Phi-2 Insights & Evaluation

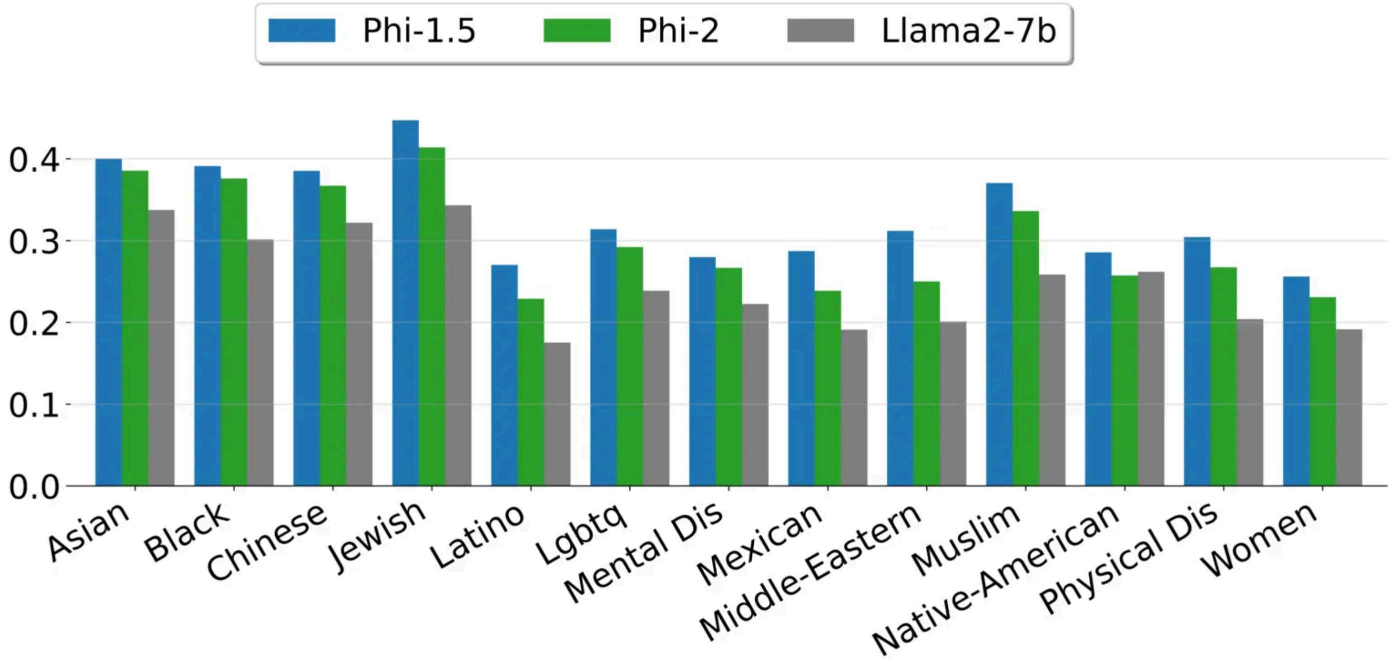
LLM researchers are keen to explore whether small language models have similar emergent capabilities as their large counterparts and if there are techniques for training that can help to achieve this.

The model is trained on "textbook-quality" data (1.4 trillion tokens with multiple passes) including synthetic datasets that help teach the model common sense reasoning and general knowledge. The data is augmented with educational and highquality web content. Phi-2 took 14 days to train on 96 A100 GPUs. No additional RLHF or instruction tuning has been applied.

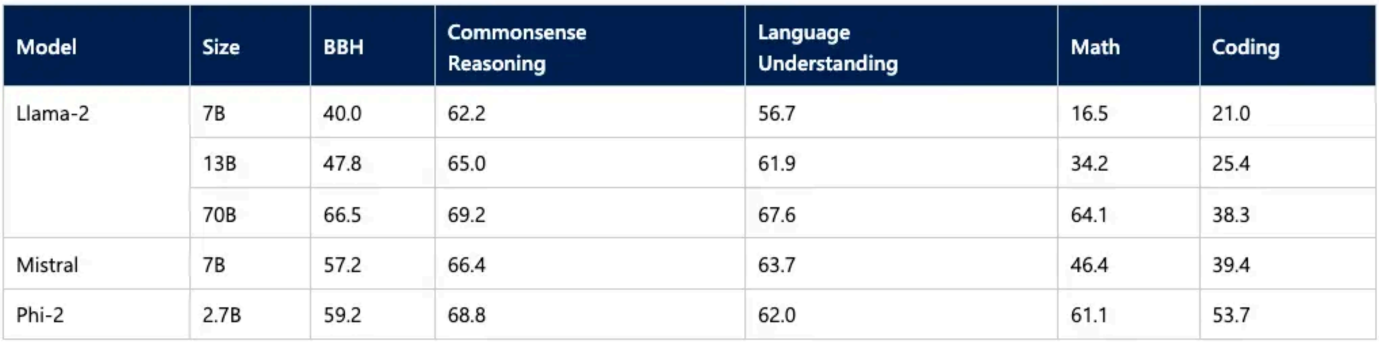
Phi-1.5 knowledge is transferred to Phi-2 which helps in model convergence and performance boost across several benchmarks. The figure below demonstrates the performance comparison between Phi-2 (2.7B) and Phi-1.5 (1.3B) on common sense reasoning, math reasoning, code generation, and other language understanding benchmarks. It's important to note that all tasks are evaluated using 0-shot with the exception of BBH and MMLU which use 3-shot CoT and 5-shot, respectively.



While the model wasn't aligned with any special technique like RLHF, it's reported to be safer in terms of toxicity and bias compared to the aligned open-source Llama2-7b. The authors attribute this to data curation efforts.

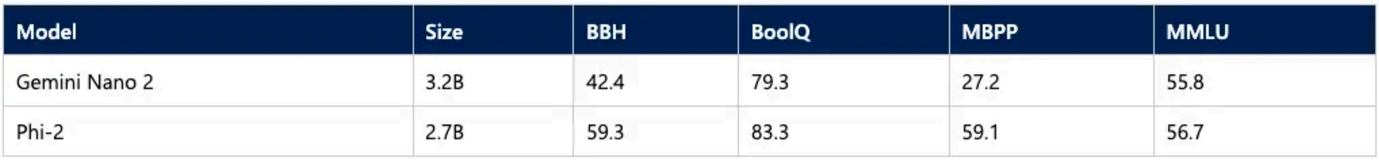


As shown in the figure below, Phi-2 outperforms Mistral 7B and Llama 2 (13B) on various benchmarks. Phi-2 even outperforms the Llama-2-70B model on multi-step reasoning. Phi-2 also outperforms Google's Gemini Nano 2.



Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.



Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Table 2. Comparison between Phi-2 and Gemini Nano 2 Model on Gemini's reported benchmarks.

Below, we provide several prompt examples that demonstrate the capabilities of the Phi-2 model on several tasks.

Physics Problem Solving

The example below demonstrates the ability of Phi-2 to solve physics word problem:

Error Identification

The example below demonstrates the ability of Phi-2 to identify the mistake in a student's calculation to a physics word problem:

Phi-2 Usage

According to the model page, Phi-2 can be prompted using a QA format, a chat format, and the code format. Below we demonstrated how to effectively use these prompt templates using different scenarios.

QA Format

QA format is useful for scenarios where you are asking the model a question and want a concise answer in return.

You can use the following prompt template:

```
Instruct: {{prompt}}
Output:
```

Here is an example:

Prompt:

```
Instruct: What is the difference between data and information?
Output:
```

Output:

```
Great question! Data refers to raw facts, statistics, or numbers that have not been analyzed
or processed into any meaningful context. On the other hand, information is data that has
been organized, analyzed, and interpreted in a way that it can be used by humans to make
informed decisions or take actions. In other words, data is raw material, while information
is processed and useful knowledge.
```

Chat Format

The chat format used by Phi-2 is a standard format and here is a simple example:

Prompt:

```
Human: Hello, who are you?
AI: Greetings! I am an AI research assistant. How can I help you today?
Human: Can you tell me about the creation of black holes?
AI:
```

Output:

```
Black holes form when a massive star runs out of fuel and collapses under its own
gravitational pull. The resulting object is so dense that it distorts space-time around it,
causing anything that comes too close to be pulled in by the immense gravity. This includes
light, which cannot...
```

Code Format

Below is a code generation prompt template that provides the name of the function to the model:

Prompt:

```
def multiply(a,b):\n
```

Keep in mind that the model has been trained on limited Python code examples and libraries so it's quite limited in its code generation capabilities and language support.

Phi-2 Limitations

Below is a summary of limitation of Phi-2, as reported by the authors:

- Similar to other models, Phi-2 may generate inaccurate code and statements.
- Phi-2 is not instruction tuned as other models and might struggle to follow instructions.
- The training consists of standard English; therefore, the model may struggle with slang and fail to comprehend instructions from other languages.
- Phi-2 may also produce societal biases and toxic content.
- Phi-2 is not tuned and tends to generate verbose responses, sometimes even producing irrelevant extra text. The authors suggest that this is probably due to the nature of the training dataset which is primarily textbooks.

Figure Sources: Microsoft Research

References

- Textbooks Are All You Need
- Phi-1.5

Last updated on September 19, 2024

Copyright © 2024 DAIR.AI