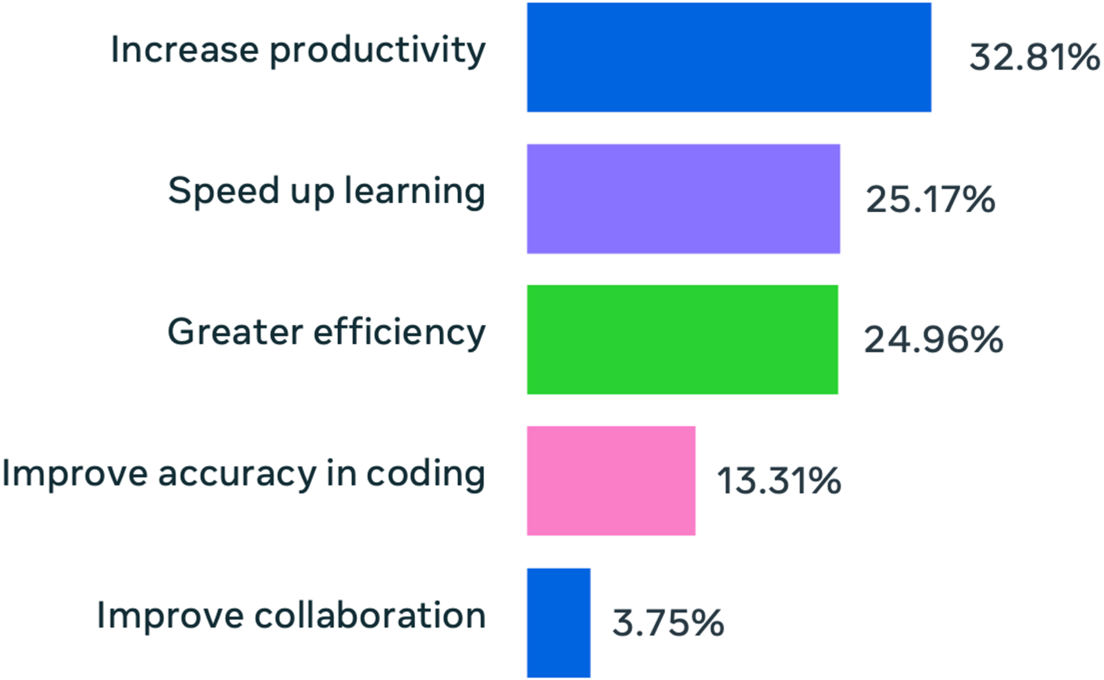
Code Llama: Open Foundation Models for Code

28m+ developers in the world



Source: Stack Overflow Survey 2023.

Code generation

- Program synthesis is not new, *including DL based*, e.g. (DeepCoder, Balog et al., 2016), (Bošnjak et al., 2017).
- Renewed interests with LLMs (e.g. Codex, 2021)

Tasks:

- Code completion
- Program synthesis from input/output pairs
- Linting
- Typing
- Bug finding
- Tests generation
- Translation

Existing models

Closed models

- Running on GPUs on servers
- Inaccessible model weights

Open models (Llama, StarCoder)

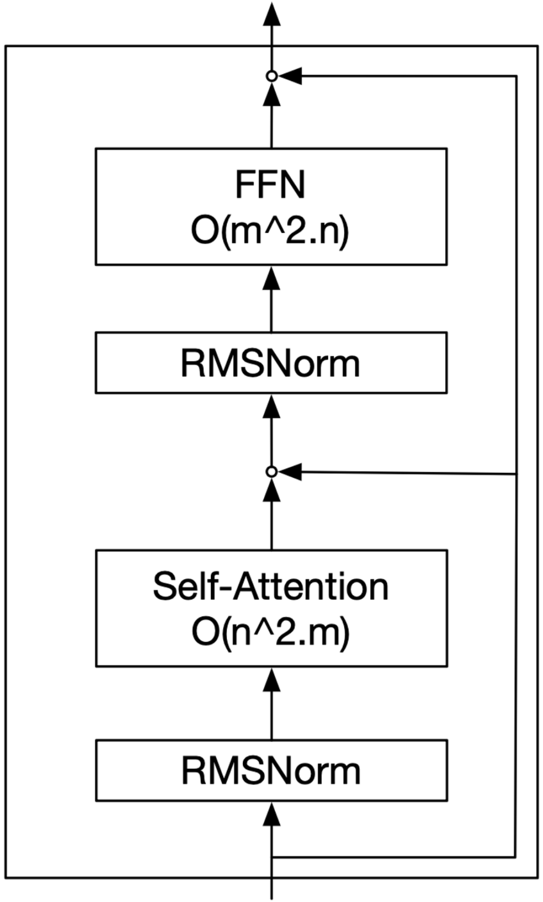
- Can be finetuned for particular language/ codebases
- Can run locally, with no internet connection
- Benefit from community improvements
- The models are **free** to use, no license fee

LLM 101

- P(next token | all previous tokens)
    - For all tokens
    - For **lots of text**
- `<bos> the cat sat on  the mat <eos>`

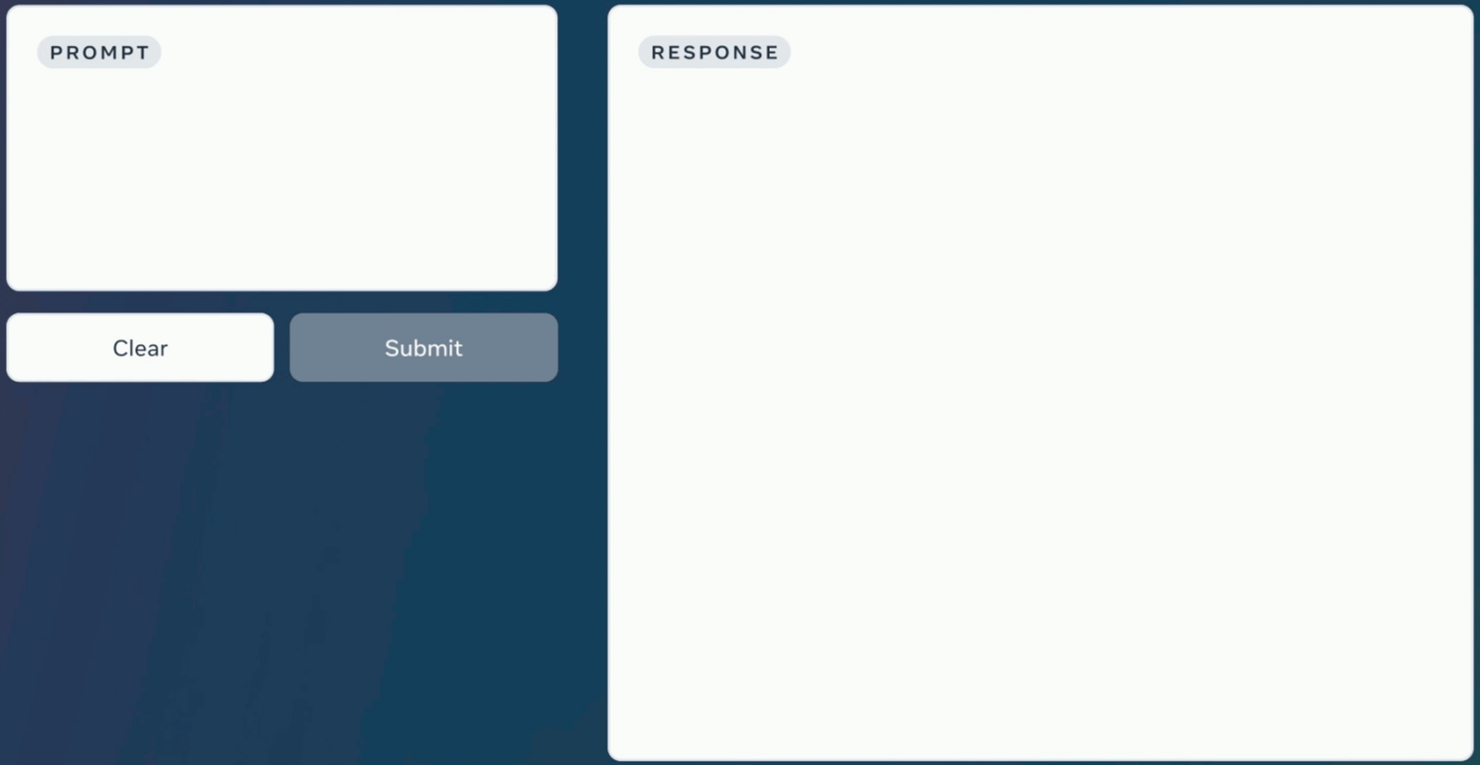
  Teacher forcing:

```
<bos> ???
<bos> the ???
<bos> the cat ???
<bos> the cat sat ???
<bos> the cat sat on  ???
<bos> the cat sat on  the ???
<bos> the cat sat on  the mat ???
```

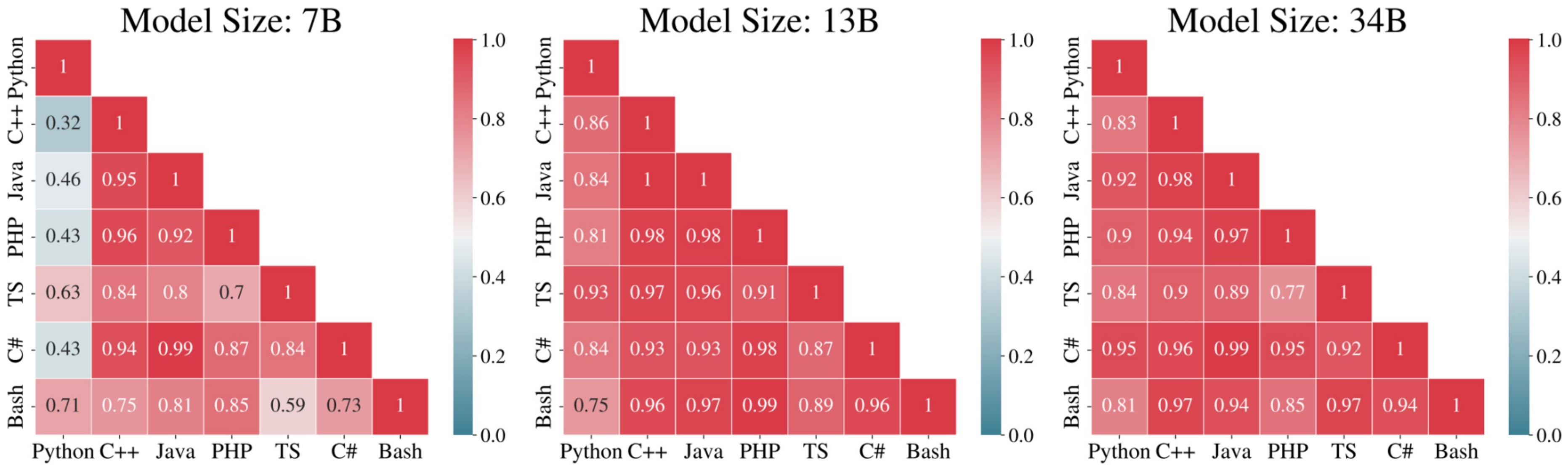


Meta AI

Code Llama



Generating Code Llama's paper figures with Code Llama



Model Size: 7B

	Python	C++	Java	PHP	TS	C#	Bash
Python	1						
C++	0.32	1					
Java	0.46	0.95	1				
PHP	0.43	0.96	0.92	1			
TS	0.63	0.84	0.8	0.7	1		
C#	0.43	0.94	0.99	0.87	0.84	1	
Bash	0.71	0.75	0.81	0.85	0.59	0.73	1

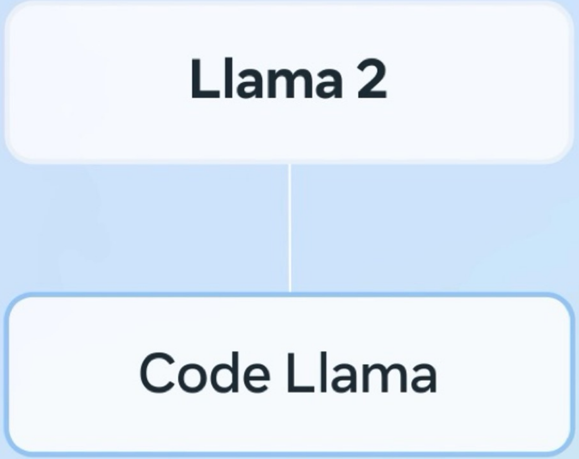
Model Size: 13B

	Python	C++	Java	PHP	TS	C#	Bash
Python	1						
C++	0.86	1					
Java	0.84	1	1				
PHP	0.81	0.98	0.98	1			
TS	0.93	0.97	0.96	0.91	1		
C#	0.84	0.93	0.93	0.98	0.87	1	
Bash	0.75	0.96	0.97	0.99	0.89	0.96	1

Model Size: 34B

	Python	C++	Java	PHP	TS	C#	Bash
Python	1						
C++	0.83	1					
Java	0.92	0.98	1				
PHP	0.9	0.94	0.97	1			
TS	0.84	0.9	0.89	0.77	1		
C#	0.95	0.96	0.99	0.95	0.92	1	
Bash	0.81	0.97	0.94	0.85	0.97	0.94	1

Figure 3: **Correlations between Languages.** Correlation scores between the Python, C++, Java, PHP, C#, TypeScript (TS), and Bash, reported for different model sizes. The code for this figure was generated by Code Llama - Instruct, the prompt and code can be seen in Figure 22.



Llama 2

Code Llama

Code Llama

500B tokens

Mostly programming languages

Code Llama - Python

100B extra tokens

Python code generation

Code Llama - I

lama - Python

0B extra tokens

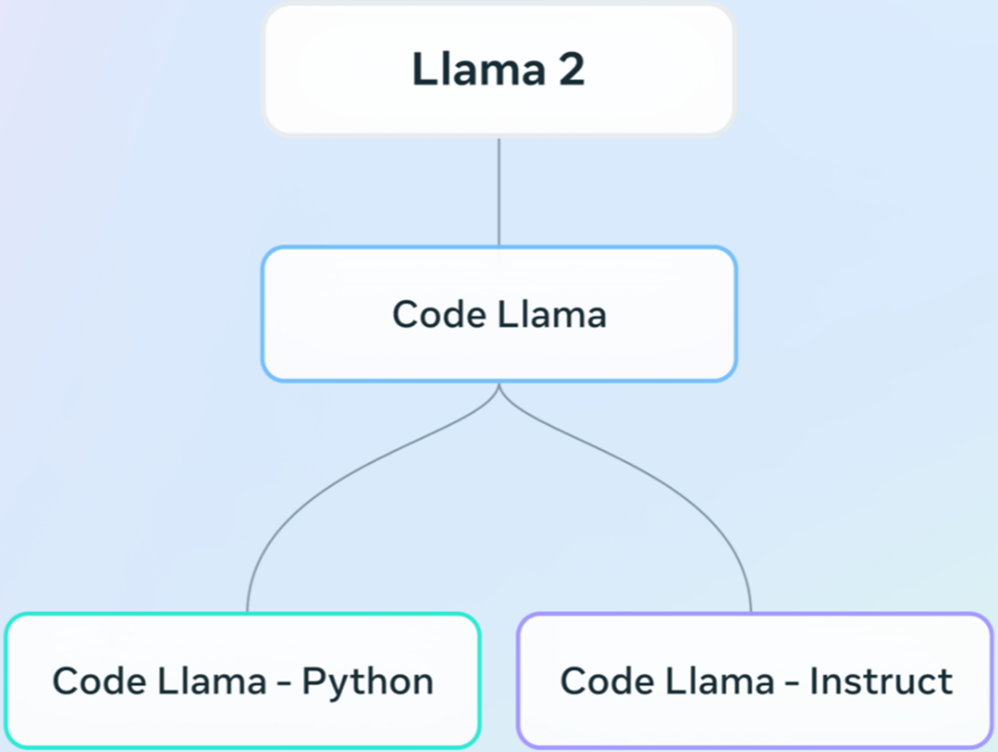
n code generation

Code Llama - Instruct

5B extra tokens

Follows instructions

Answers questions



Llama 2

Code Llama

Code Llama - Python

Code Llama - Instruct

7B

13B

34B

70B

7B

13B

34B

70B

7B

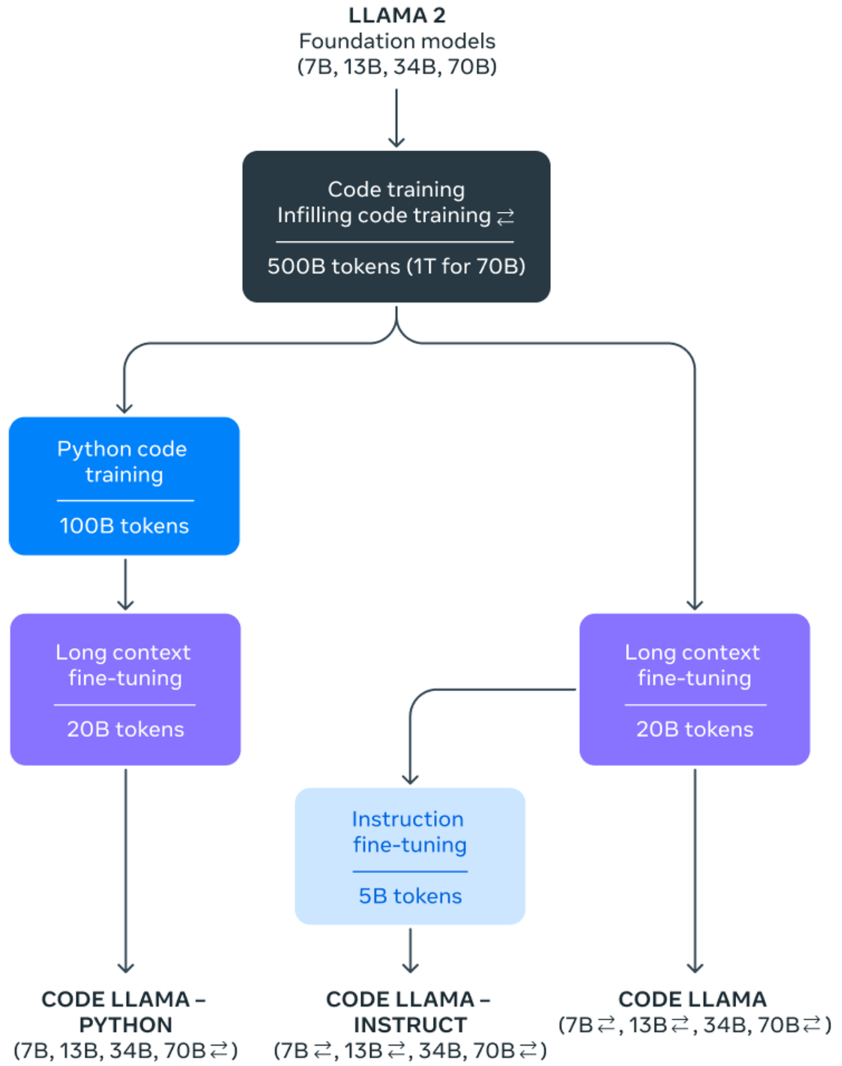
13B

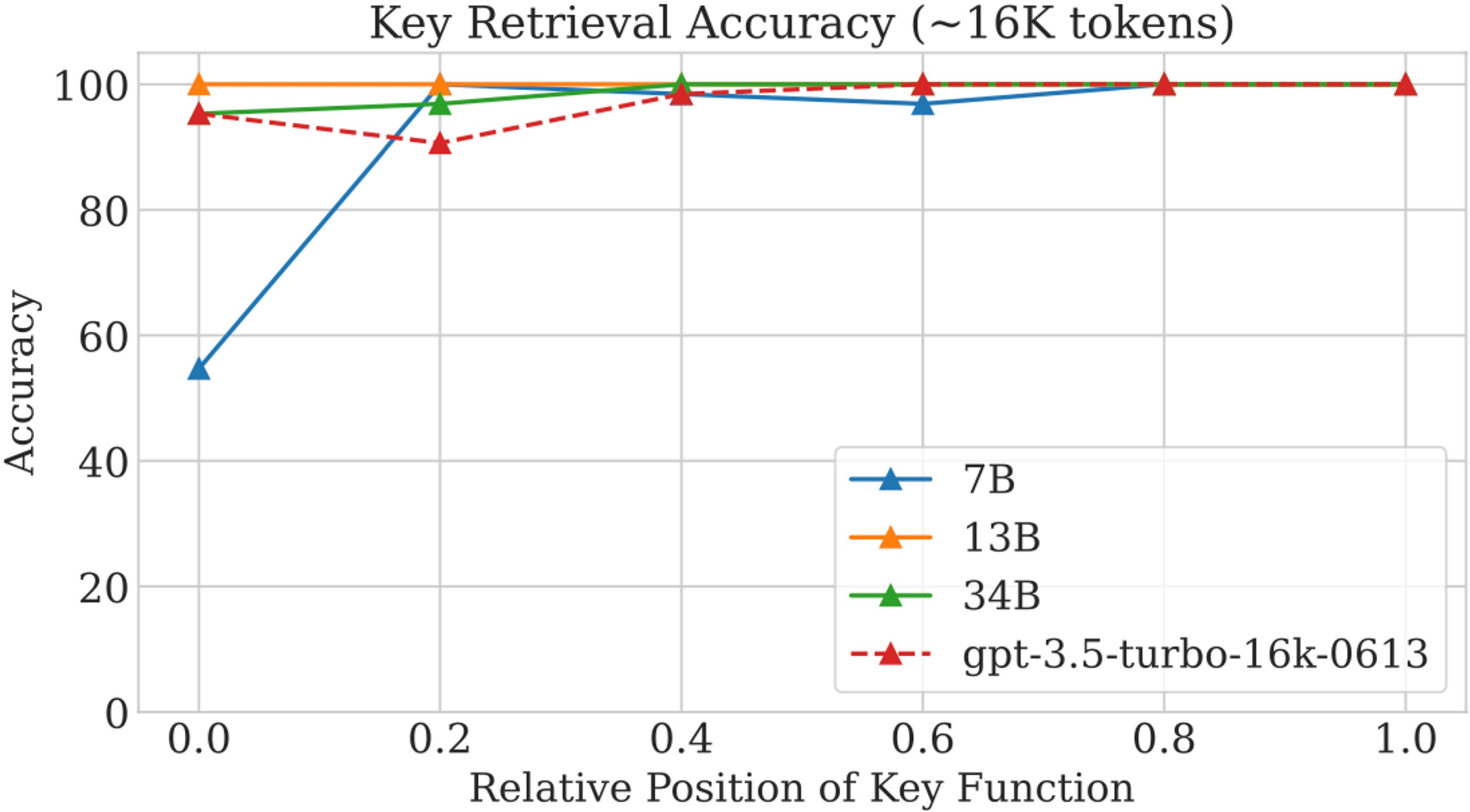
34B

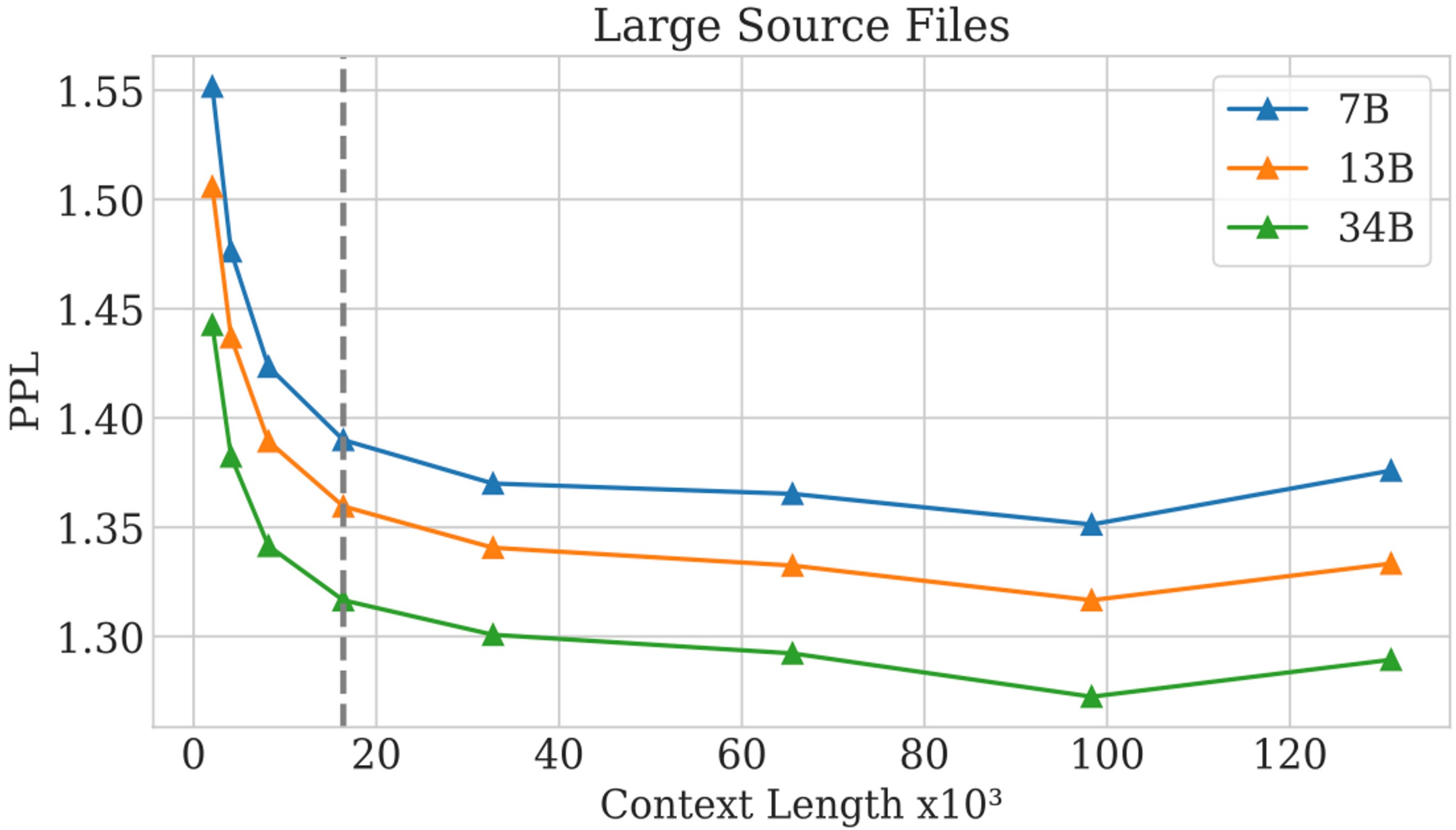
70B

Long context

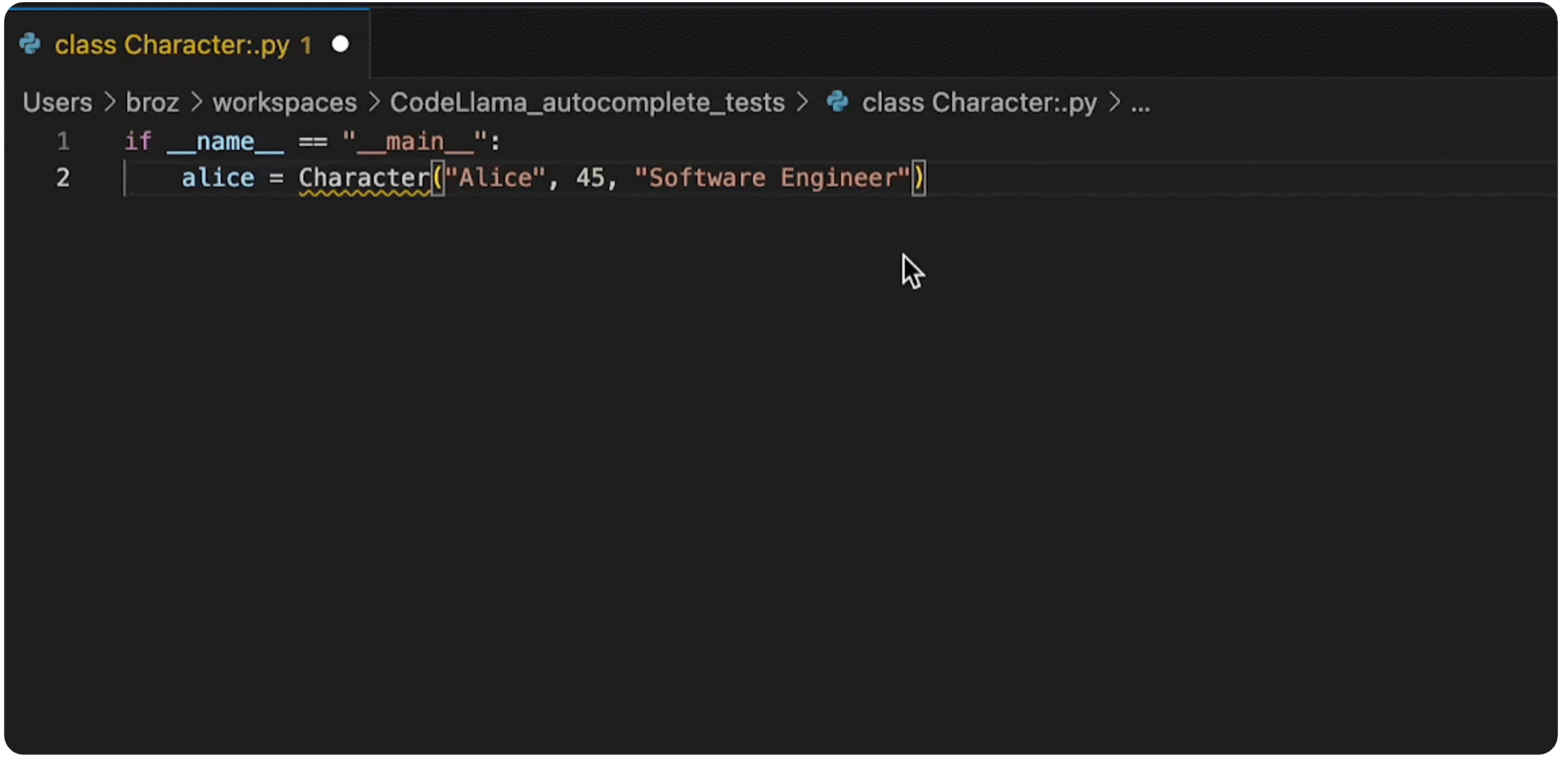
- ~20B tokens fine-tuning
- Trained with up 16k tokens
- Supports up to 100k tokens = 8k lines of code




Key Retrieval Accuracy (~16K tokens)
Accuracy
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
Relative Position of Key Function
7B
13B
34B
gpt-3.5-turbo-16k-0613


Large Source Files
PPL
1.55
1.50
1.45
1.40
1.35
1.30
Context Length x10³
0
20
40
60
80
100
120
7B
13B
34B

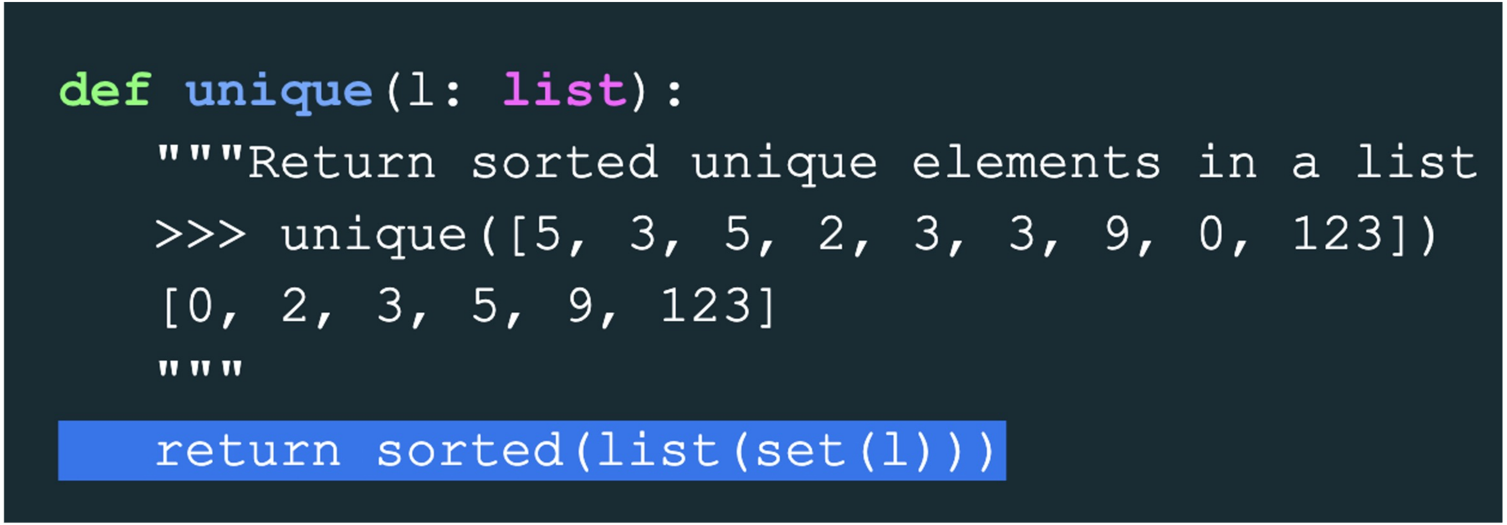
Fill-in-the-middle (FIM)



class Character:.py 1
Users > broz > workspaces > CodeLlama_autocomplete_tests > class Character:.py > ...

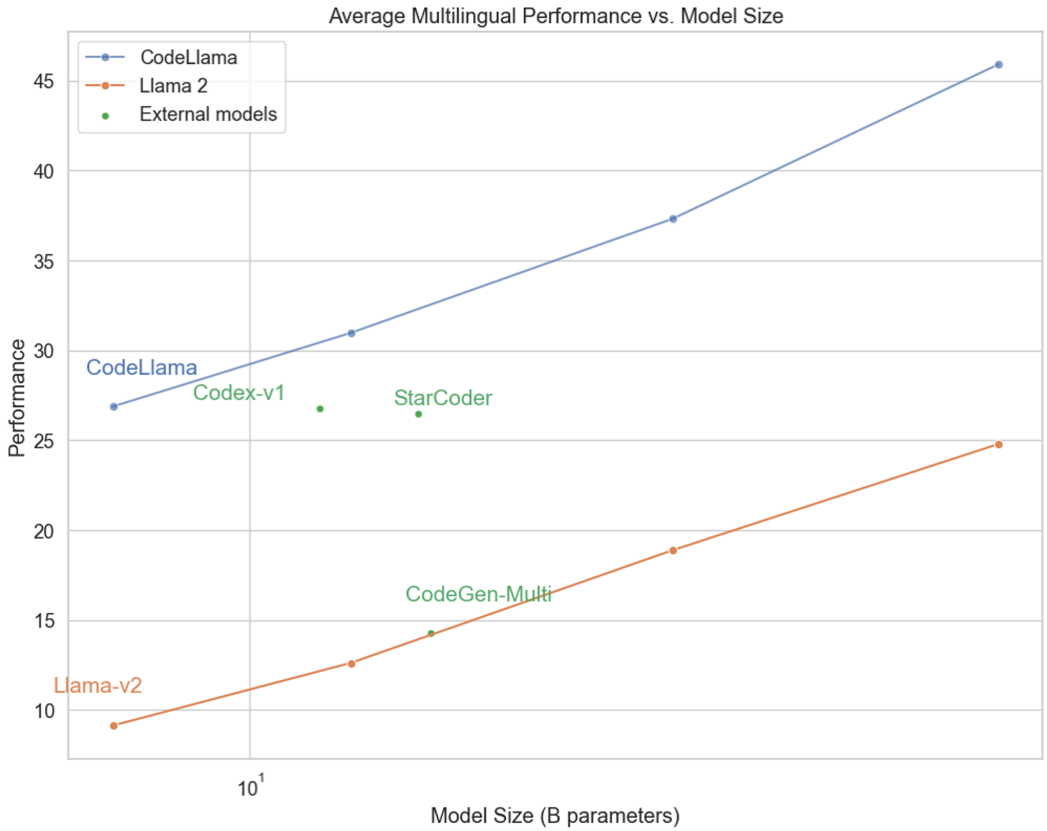
```python
if __name__ == "__main__":
    alice = Character("Alice", 45, "Software Engineer")
```

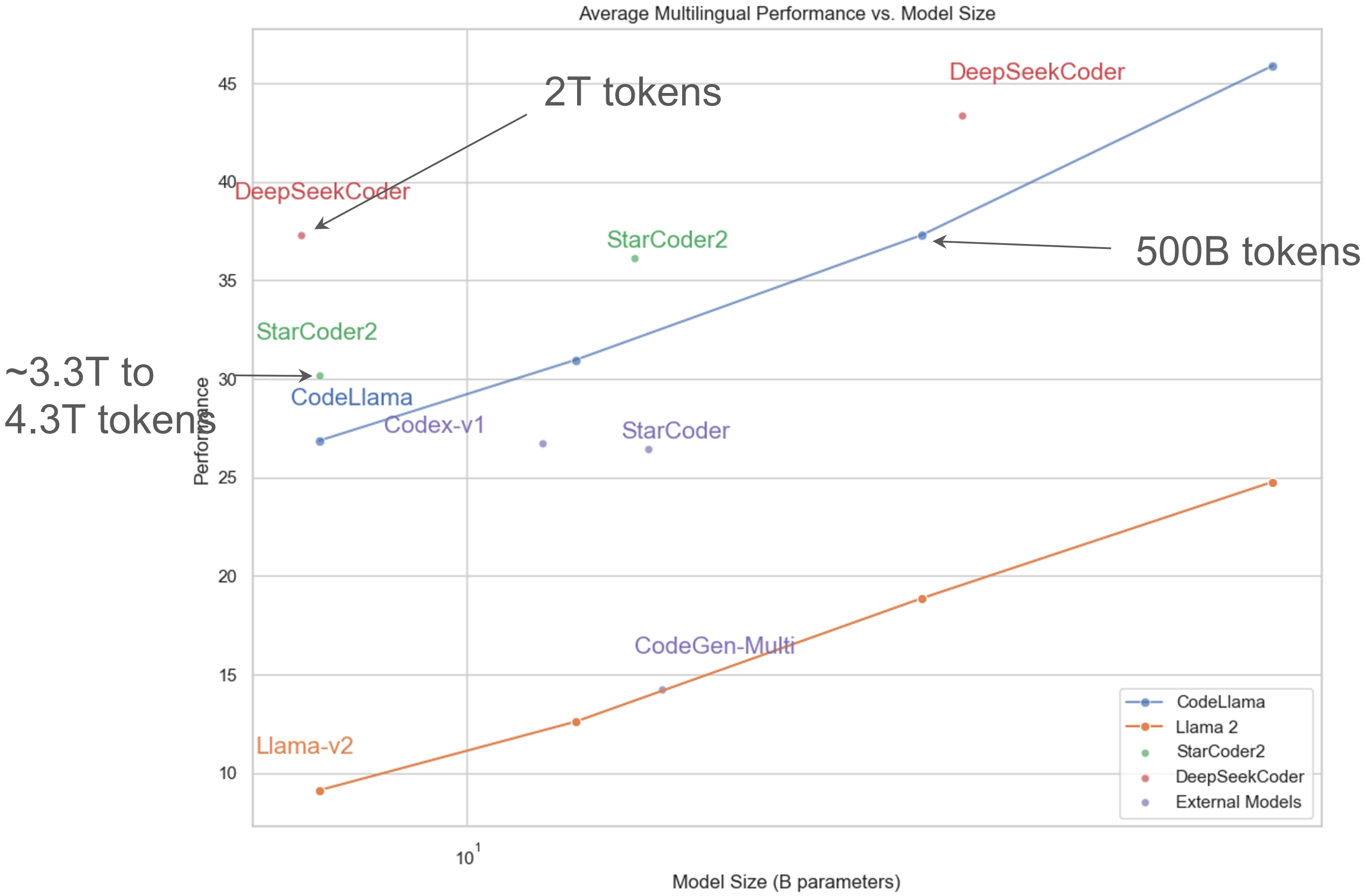
HumanEval example



```python
def unique(l: list):
    """Return sorted unique elements in a list
    >>> unique([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [0, 2, 3, 5, 9, 123]
    """
    return sorted(list(set(l)))
```

Code Llama 7B outperforms Llama 2 70B on multilingual coding benchmarks

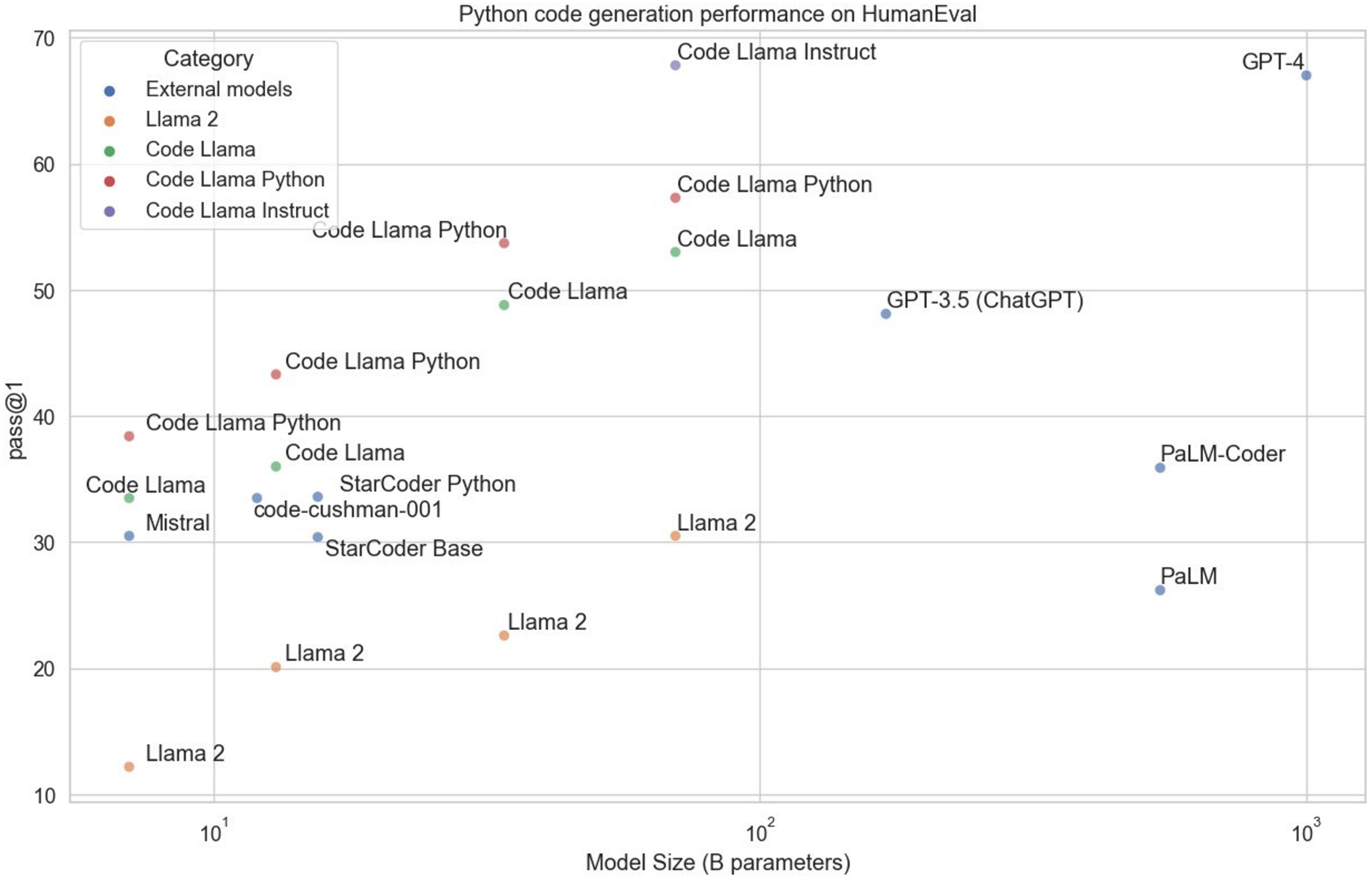


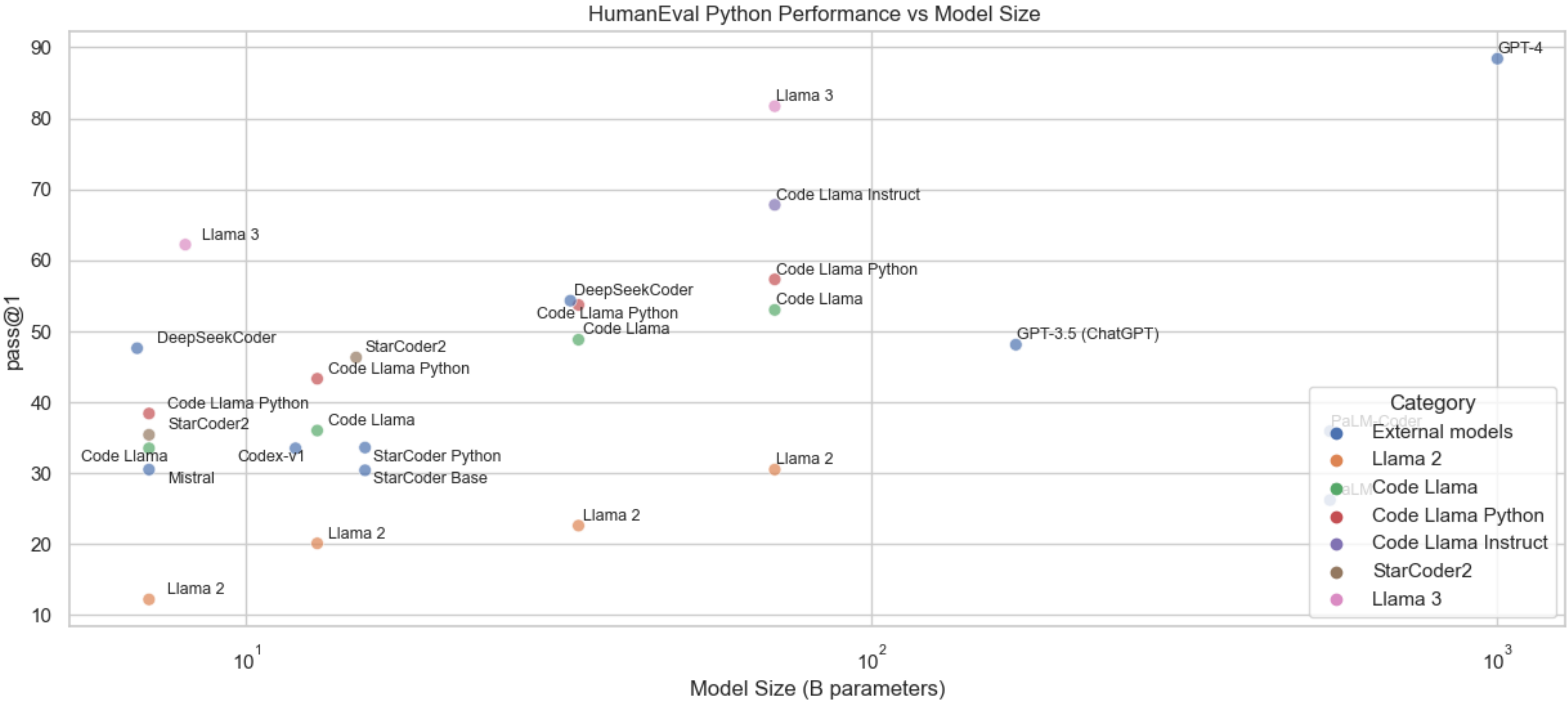


2T tokens

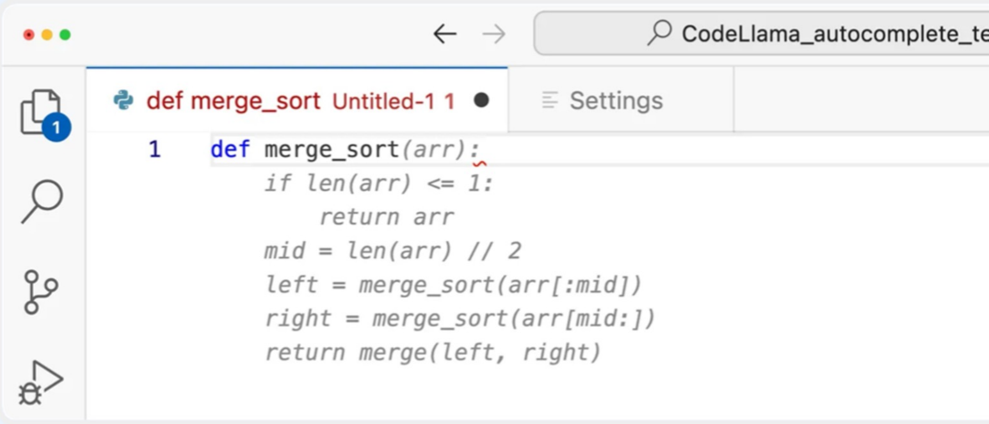
500B tokens

~3.3T to 4.3T tokens


Python code generation performance on HumanEval
Category
● External models
● Llama 2
● Code Llama
● Code Llama Python
● Code Llama Instruct
Code Llama Instruct
GPT-4
Code Llama Python
Code Llama Python
Code Llama
Code Llama
GPT-3.5 (ChatGPT)
Code Llama Python
Code Llama Python
Code Llama
StarCoder Python
Code Llama
code-cushman-001
Mistral
Llama 2
StarCoder Base
PaLM-Coder
PaLM
Llama 2
Llama 2
Llama 2
pass@1
70
60
50
40
30
20
10
10^1
10^2
10^3
Model Size (B parameters)

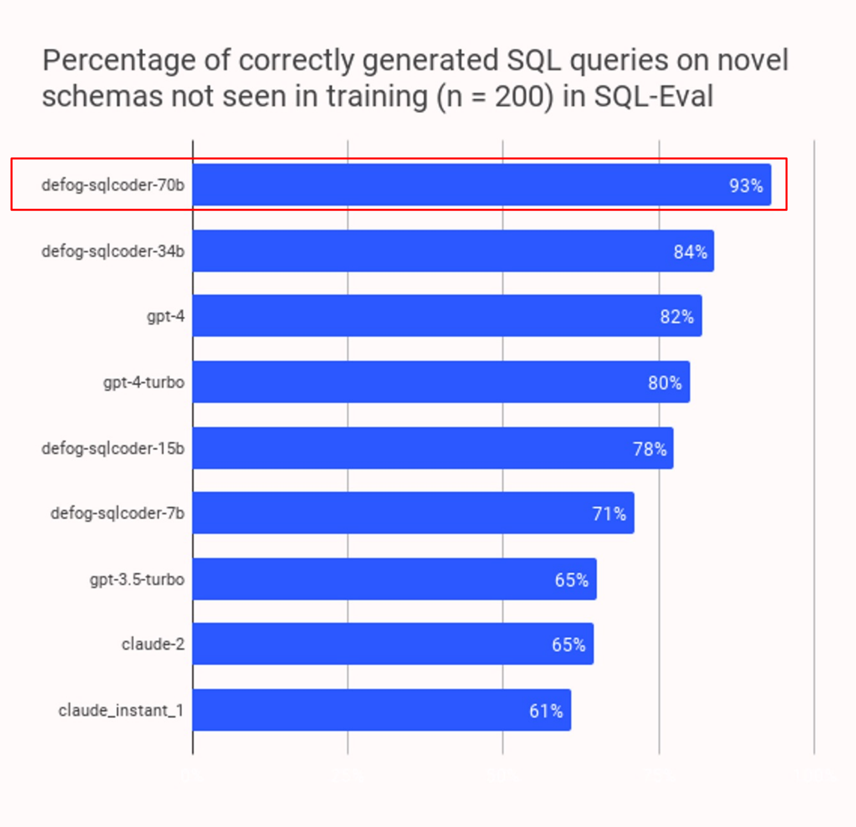

HumanEval Python Performance vs Model Size
pass@1
90
80
70
60
50
40
30
20
10
Model Size (B parameters)
10^1
10^2
10^3
GPT-4
Llama 3
Code Llama Instruct
Llama 3
Code Llama Python
DeepSeekCoder
Code Llama
Code Llama Python
Code Llama
GPT-3.5 (ChatGPT)
DeepSeekCoder
StarCoder2
Code Llama Python
Code Llama Python
StarCoder2
Code Llama
PaLM-Coder
Code Llama
Codex-v1
StarCoder Python
Llama 2
Mistral
StarCoder Base
aLM
Llama 2
Llama 2
Llama 2
Category
● External models
● Llama 2
● Code Llama
● Code Llama Python
● Code Llama Instruct
● StarCoder2
● Llama 3

Hugging Face integration with VSCode

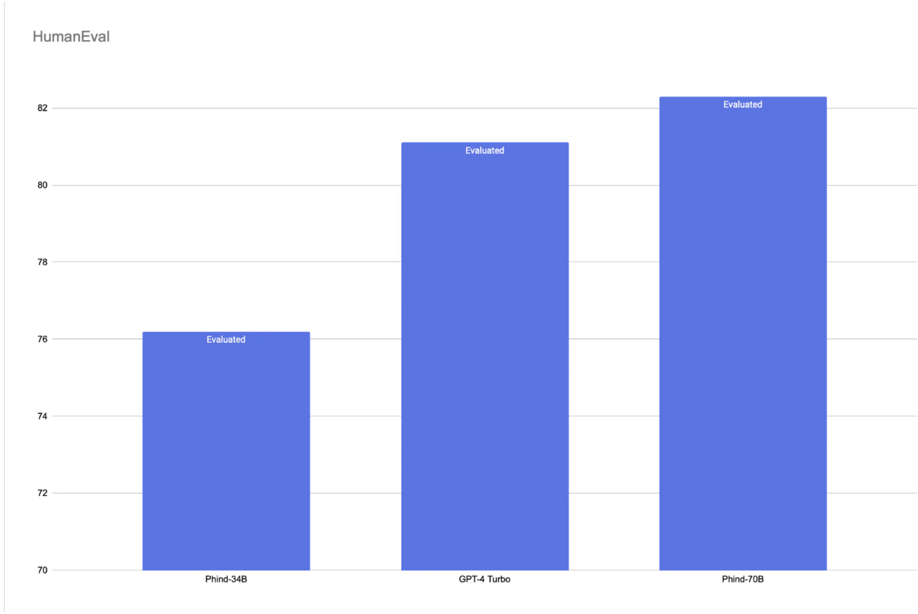


```
def merge_sort(arr):
    if len(arr) <= 1:
        return arr
    mid = len(arr) // 2
    left = merge_sort(arr[:mid])
    right = merge_sort(arr[mid:])
    return merge(left, right)
```

Built on Code Llama, 2 examples:



Percentage of correctly generated SQL queries on novel schemas not seen in training (n = 200) in SQL-Eval



HumanEval

https://twitter.com/rishdotblog/status/1752329471867371659 20k instructions || https://www.phind.com/blog/introducing-phind-70b 50B tokens

Write a simple version of pong using pygame.

Get started with Code Llama

- Ollama https://ollama.com/library/codellama
- HuggingFace https://huggingface.co/codellama/
- Perplexity AI chat https://labs.perplexity.ai/
- Our inference GitHub repository https://github.com/facebookresearch/codellama

Questions ?