Project Website

Paper

Code

InfiBench: Evaluating the Question-Answering Capabilities of Code LLMs

Linyi Li, Shijie Geng, Zhenwen Li, Yibo He, Hao Yu,
Ziyue Hua, Guanghan Ning, Siwei Wang, Tao Xie, Hongxia Yang

Simon Fraser University Rutgers University Peking University
The Hong Kong Polytechnic University ByteDance Inc

NEURAL INFORMATION PROCESSING SYSTEMS

Benchmarks for Code LLMs

- Code LLMs:
  - Trained on code-domain data
  - Strong at coding, reasoning, UI interaction, …
- Various code benchmarks evaluate code LLMs

Limitations & Design Goals

Limitations:

- Only focus on code generation
- Most are derived datasets
  - Few from independent data source
- Benchmarks are saturating
- May be contaminated

Design Goals:

- Code-related Question-Answering
- Independent data source
- Non-saturating

SFU

Sourcing

stackoverflow

> 20M Question Posts

Filtering

17,402 Question Posts

Area Identifying & Sampling

702 Questions Posts

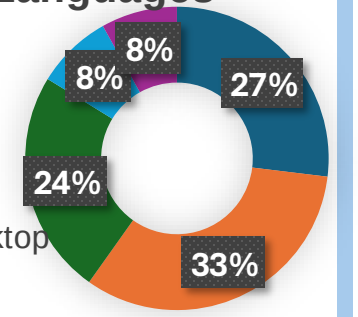
Continuously Expanding…

Labelling

Domain Experts

- Select
- Label
- Paraphrase
- Post-Filtering

Covering 5 Major Areas & 15 Languages



Evaluation

- Model-Free & Question-Specific Evaluation Criteria
- HuggingFace-Compatible Automatic Evaluation Framework

Question Example	*Metric*	
`# lang = PHP` How can I increase the laravel 8 dd() limitations? ,,,	Keywords Matching	- [0.5pt] match "`override`" - [0.5pt] match "`dumper/dump`"

`# lang = dart`
What is the right way to disable back button … repeat the following paragraph with [blank] filled:
`Use [blank] instead of [blank].`
	Blank Filling	`Use pushAndRemoveUntil()`
		[0.5pt] `instead of pop/pop()`
		[0.5pt].

`# lang = C++`
Complete the function ```` ```vector <pair<int, int>> ```` The function create …
	Unit Testing	`int main(){`
		`vector<int> origin = …`
		`// assert …`
		`…`

`# lang = Java` How to enable Dev Tools project on IntelliJ 2021.2 …	Dialogue Similarity	Linearly scale ([0.30, 0.51] -> [0pt, 1pt]) rougeLSum score with reference answer

Leaderboard

Comprehensive Benchmark of **100+** Large Language Models Leading to Several Findings

Key Takeaways

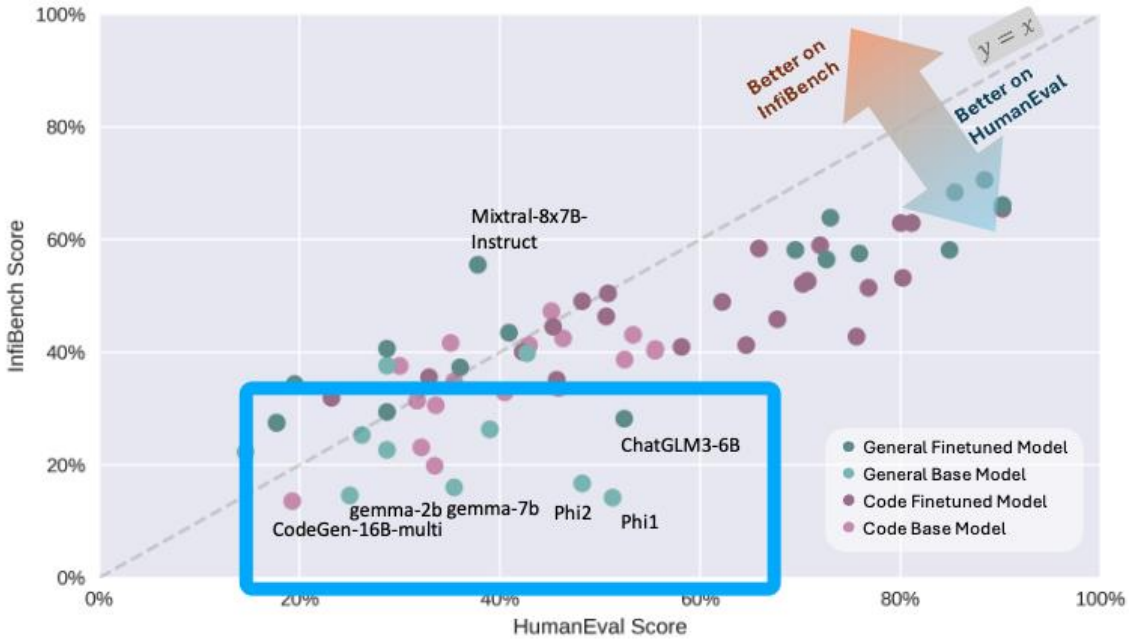
- GPT-4 far from perfect, open-source models close but not exceed GPT-4 yet
- Among models of same sizes, their performances vary
- Hard problems generalize
- Instruction finetuning is important

General Base , General Finetuned , Code Base , Code Finetuned

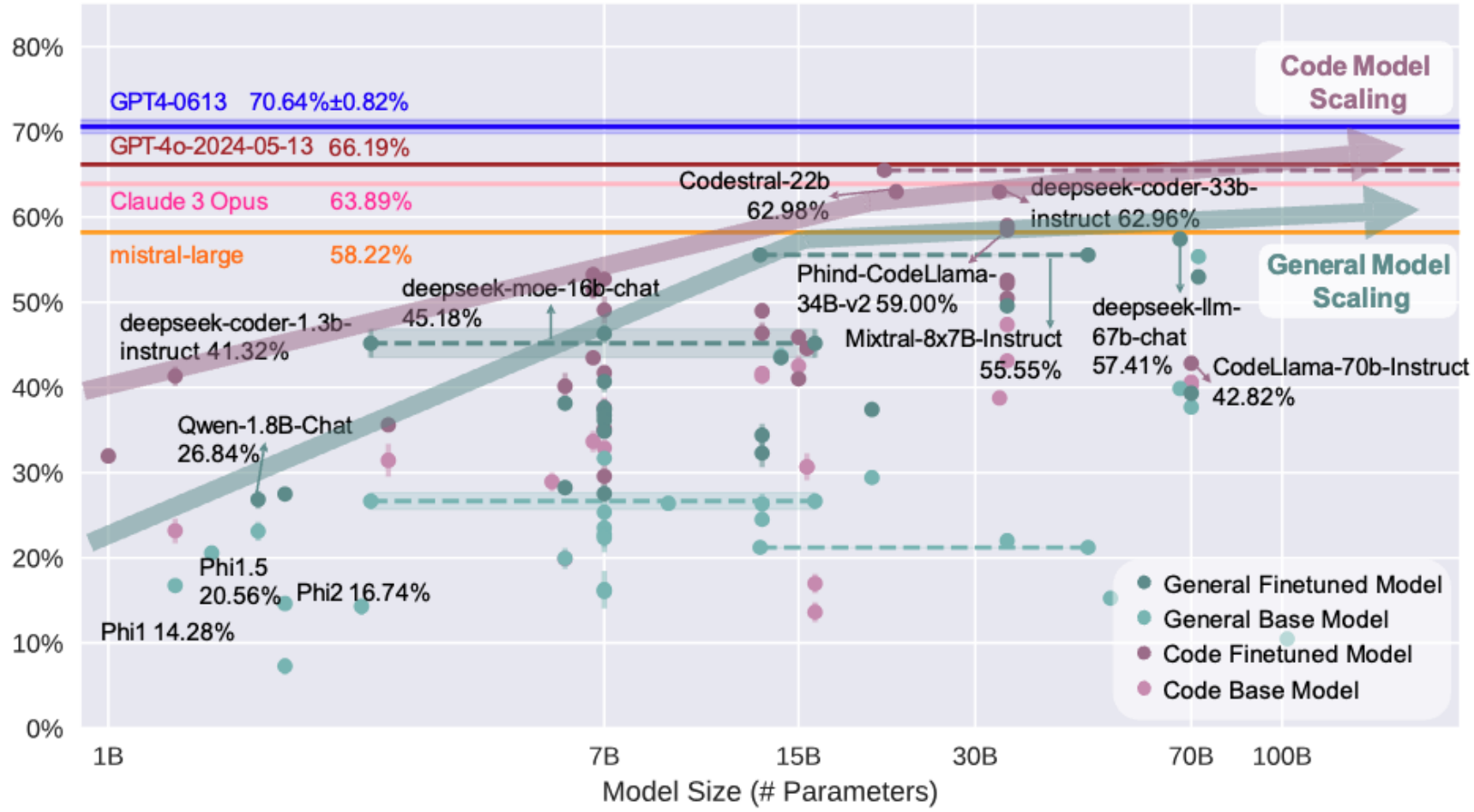
	Family	Best Model Name	Size	InfiBench Score
1	GPT-4	GPT-4-0613	?	70.64% ± 0.82%
2	DeepSeek Coder	deepSeek-coder-V2-instruct	236B / 21B	65.49%
3	Claude 3	Claude 3 Opus	?	63.89%
4	Mistral Open	Codestral-22b	22B	62.98% ± 0.56%
5	Phind	Phind-CodeLlama-34B-v2	34B	59.00%
6	Mistral	mistral-large	?	58.22%
7	DeepSeek LLM	deepseek-llm-67b-chat	67B	57.41%
8	GPT-3.5	GPT-3.5-turbo-0613	?	56.47% ± 1.34%
9	Qwen	Qwen-72B	72B	55.34%
10	Magicoder	Magicoder-S-CL-7B	7B	52.71% ± 0.72%
11	WizardLM	WizardCoder-Python-34B-V1.0	34B	52.59%
12	Code Llama	CodeLlama-34b-Instruct	34B	50.45%
13	01.AI	Yi-34B-Chat	34B	49.58%
14	Zephyr	Zephyr 7B beta	7B	46.31% ± 1.11%
15	StarCoder2	15B-Instruct	15B	45.89% ± 0.95%
16	DeepSeek MoE	deepseek-moe-16b-chat	16B / 2.8B	45.18% ± 1.65%
17	OctoPack	OctoCoder	15.5B	44.55% ± 0.79%
18	gemma	gemma-7b-it	7B	40.68% ± 1.23%
19	Llama 2	Llama2-70B-Chat	70B	39.30%
20	InternLM	InternLM-Chat-20B	20B	37.41% ± 0.75%
21	Baichuan2	Baichuan2-13B-Chat	13B	34.40% ± 1.34%
22	StarCoder	StarCode+	15.5B	30.67% ± 1.57%
23	CodeGen2.5	CodeGen2.5-7B-Instruct	7B	29.57% ± 1.53%
24	ChatGLM	ChatGLM3-6B	6B	28.23% ± 0.58%
25	davinci	davinci-002	?	21.25% ± 1.17%
26	Phi	Phi1.5	1.5B	20.56% ± 0.09%
27	CodeGeeX	CodeGeeX2-6B	6B	19.88% ± 0.36%
28	CodeGen2	CodeGen2-16B	16B	16.97% ± 1.15%
29	IEITYuan	Yuan2-51B-hf	51B	15.25%
30	CodeGen	CodeGen-16B-multi	16B	13.62% ± 1.18%
	Human	10 Highest-Voted Answer Posts		65.18%
		Highest-Voted Answer Post		56.28%
		Officially-Accepted Answer Post		52.90%

Key Takeaways

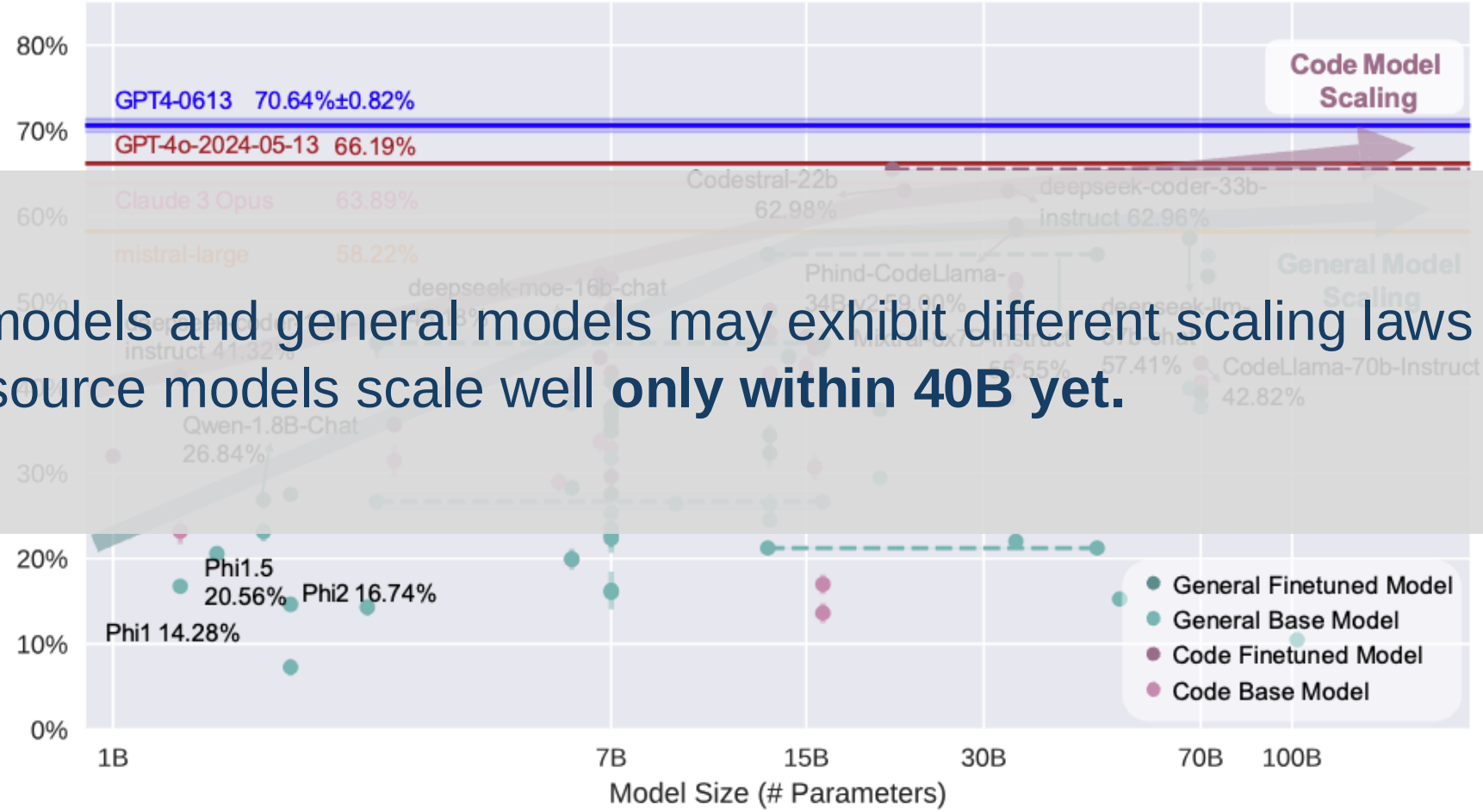
- GPT-4 far from perfect, open-source models close but not exceed GPT-4 yet
- Among models of same sizes, their performances vary
- Hard problems generalize
- Instruction finetuning is important
- Some models overly focus on code generation, ignoring other capabilities



Empirical Scaling Laws



Empirical Scaling Laws



- Code models and general models may exhibit different scaling laws
- Open-source models scale well **only within 40B yet.**