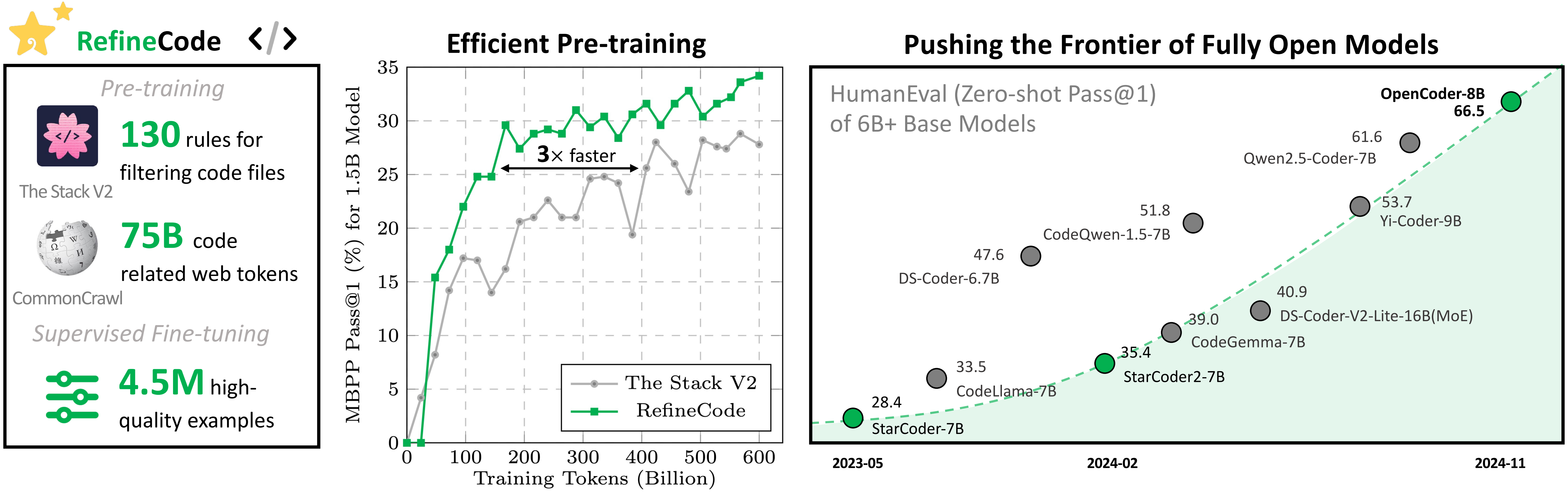
RefineCode
Pre-training
130 rules for filtering code files
The Stack V2
75B code related web tokens
CommonCrawl
Supervised Fine-tuning
4.5M high-quality examples
Efficient Pre-training
3× faster
MBPP Pass@1 (%) for 1.5B Model
Training Tokens (Billion)
0
5
10
15
20
25
30
35
0
100
200
300
400
500
600
The Stack V2
RefineCode
Pushing the Frontier of Fully Open Models
HumanEval (Zero-shot Pass@1)
of 6B+ Base Models
OpenCoder-8B
66.5
61.6
Qwen2.5-Coder-7B
53.7
Yi-Coder-9B
51.8
CodeQwen-1.5-7B
47.6
DS-Coder-6.7B
40.9
DS-Coder-V2-Lite-16B(MoE)
39.0
CodeGemma-7B
35.4
StarCoder2-7B
33.5
CodeLlama-7B
28.4
StarCoder-7B
2023-05
2024-02
2024-11