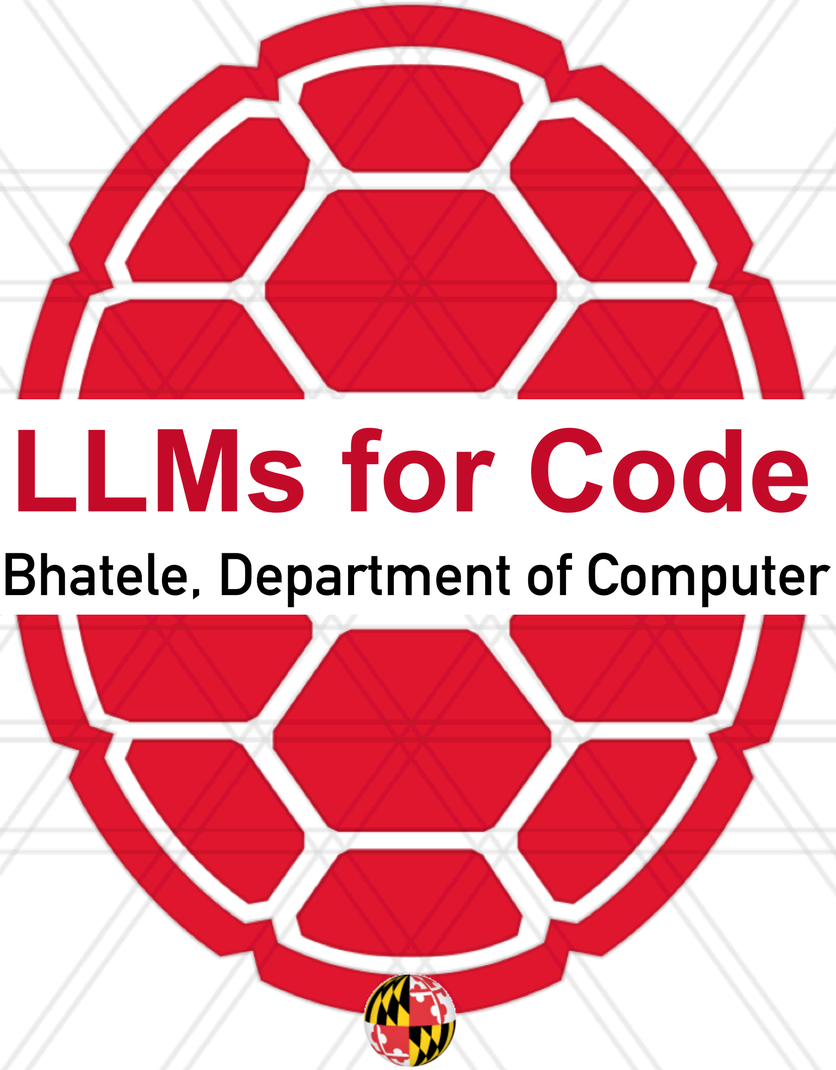
Systems for Machine Learning (CMSC828G)



LLMs for Code

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Anecdotal uses of AI in education / research

- Creating exams from pdfs of books and slides
- Creating summaries/lecture notes from videos
- Virtual course assistants for students
- Course assignments
  - Students using AI for writing code and essays

LLMs for code and software development

- Code LLMs: fine-tuned using code data
- Performance is evaluated on the ability to generate code
- Useful for a variety of code development tasks:
  - Code completion (generation or synthesis)
  - Summarization
  - Debugging / fixing bugs
  - Translation

Model	Organization	Global Average	Reasoning Average	Coding Average	Agentic Coding Average	Mathematics Average
GPT-5.2 Codex	OpenAI	75.44	77.71	83.62	51.67	88.77
Claude 4.7 Opus Thinking High Effort	Anthropic	79.49	83.83	83.18	61.67	89.29
GPT-5.1 Codex Max	OpenAI	76.57	84.57	81.38	56.67	83.66
Claude 4 Sonnet	Anthropic	54.78	39.67	80.74	38.33	60.36
Claude Sonnet 4.5 Thinking	Anthropic	72.65	77.59	80.36	53.33	79.31
GPT-5.1 Low	OpenAI	62.08	59.64	80.30	40.00	68.38
Claude 4.6 Sonnet Thinking High Effort	Anthropic	77.39	86.38	79.98	56.67	86.53
Claude 4.5 Opus Thinking High Effort	Anthropic	78.37	80.09	79.65	63.33	90.39
GPT-5.3 Instant	OpenAI	60.62	63.12	78.63	28.33	72.41
Gemini 3 Flash Preview Minimal	Google	59.79	49.17	78.57	43.33	68.10
Kimi K2.6 Thinking	Moonshot AI	75.14	79.38	78.57	58.33	84.28

https://livebench.ai

Challenges in using AIforDev for parallel code

- Benchmarks are not relevant to HPC
  - Popular languages are C, C++, Fortran
  - Parallel execution models: MPI, OpenMP, CUDA, HIP, Kokkos, RAJA, etc.
- Metrics for evaluating Code LLMs are not well-suited for parallel computing
  - HPC researchers care about performance and portability in addition to correctness
- Current LLMs are not great for parallel code
- Not enough data for training/fine-tuning: low-resource languages

Synthetic data: ‘code is all you need’

Synthetic data: ‘code is all you need’

- We have exhausted all available code

Synthetic data: ‘code is all you need’

- We have exhausted all available code
- Parallel code data falls in the low-resource category

Total files	545.5 million	%
MPI	185.4k	0.034
OpenMP	152.9k	0.028
Kokkos	17.7k	0.003
Fortran	287.1k	0.052
CUDA	145.6k	0.027

The Stack dataset

Synthetic data: ‘code is all you need’

- We have exhausted all available code
- Parallel code data falls in the low-resource category
- Recent work on generating synthetic code datasets using LLMs

Total files	545.5 million	%
MPI	185.4k	0.034
OpenMP	152.9k	0.028
Kokkos	17.7k	0.003
Fortran	287.1k	0.052
CUDA	145.6k	0.027

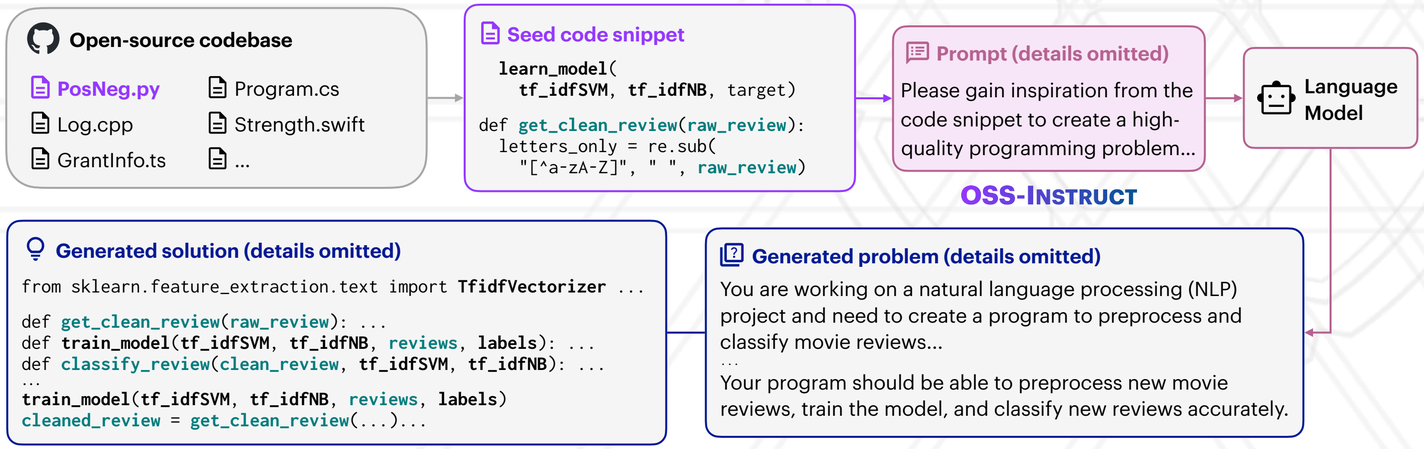
The Stack dataset

Synthetic data: ‘code is all you need’

- We have exhausted all available code
- Parallel code data falls in the low-resource category
- Recent work on generating synthetic code datasets using LLMs

Total files	545.5 million	%
MPI	185.4k	0.034
OpenMP	152.9k	0.028
Kokkos	17.7k	0.003
Fortran	287.1k	0.052
CUDA	145.6k	0.027

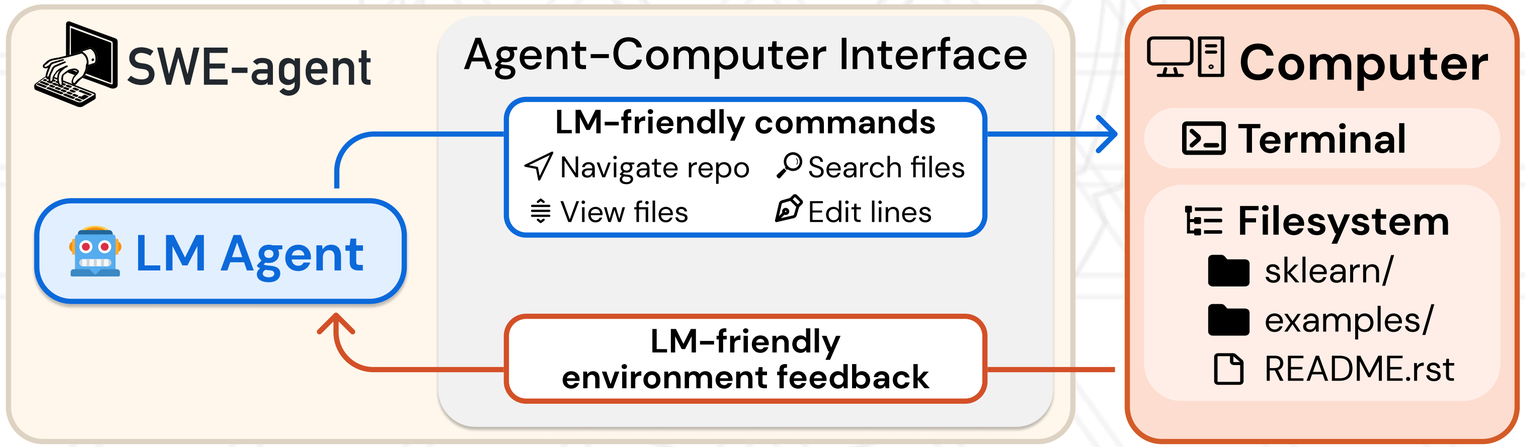
The Stack dataset



Wei et al. Magicoder: Source Code is All You Need. In Proceedings of the Forty-first International Conference on Machine Learning, ICML ’24.

SWE-agent

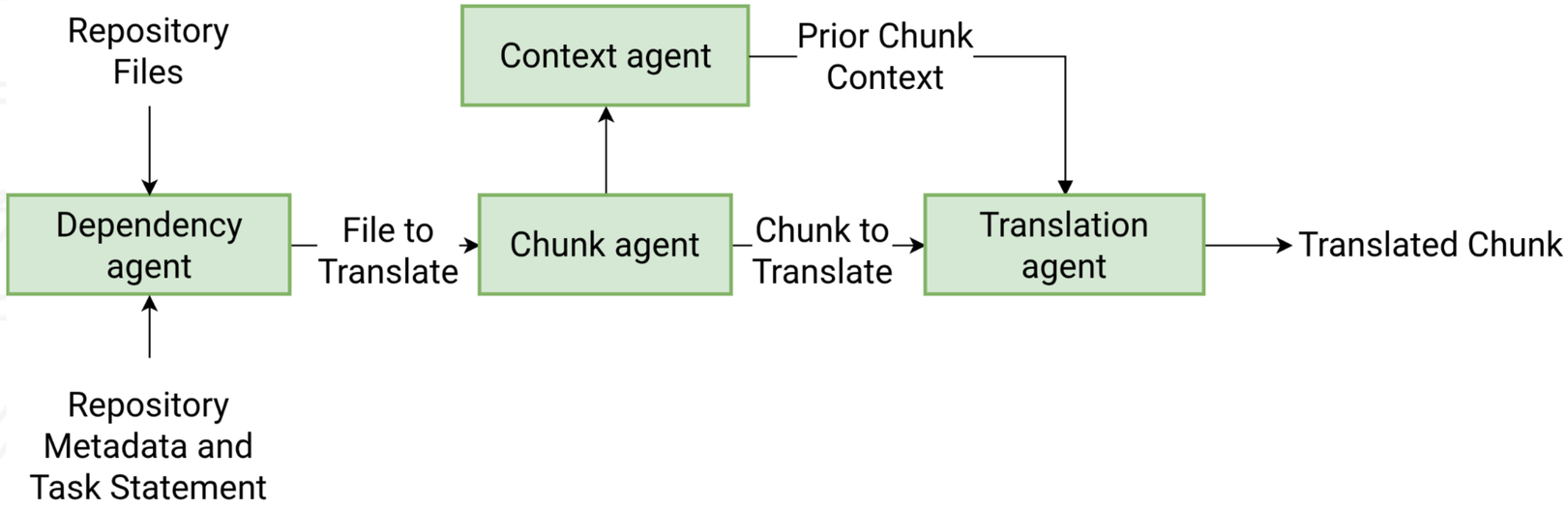
- Combines language models with an Agent-Computer interface
- SWE-agent can call tools, write tests, run code etc.
- Designed for Python primarily



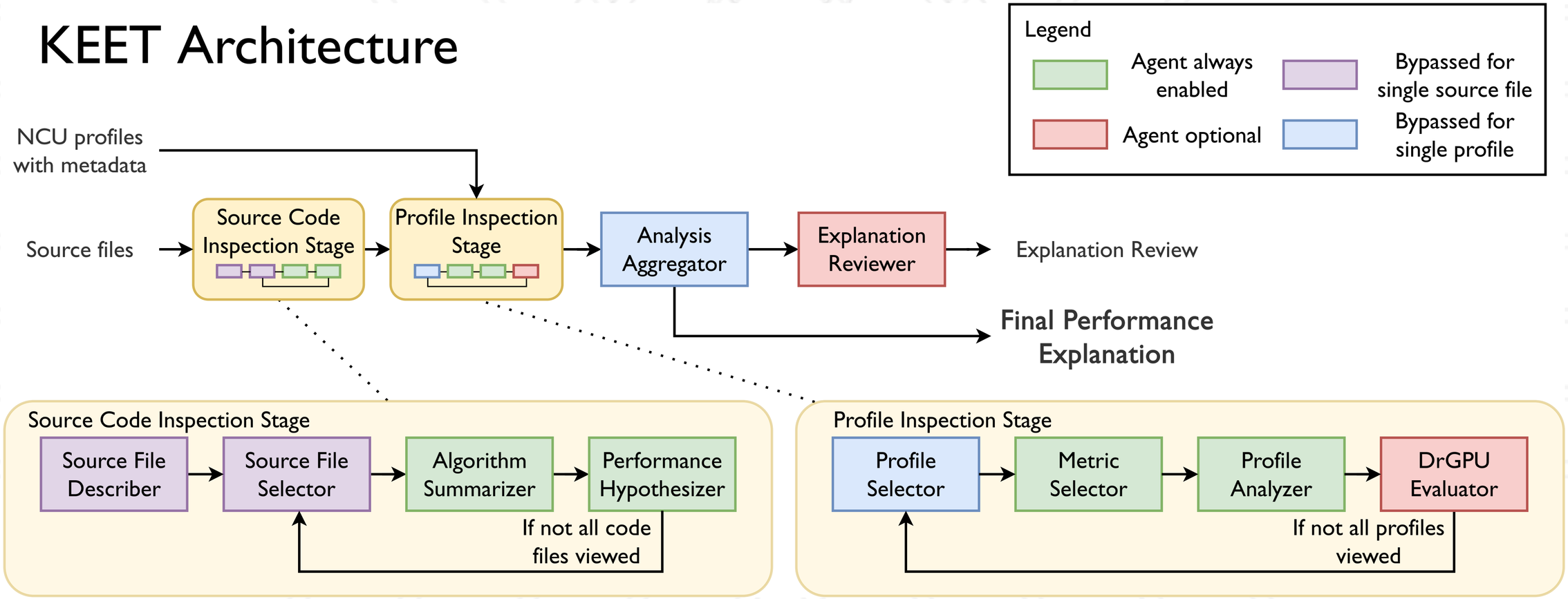
John Yang et al. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. NeurIPS 2024.

LLM agents based approach

- Naive solution or baseline: translate one file at a time in arbitrary order
  - Exceeds LLM input context limit even for very simple repositories
- Proposed solution: develop an agentic approach that divides up the translation task into smaller problems



AI agents for performance optimization



UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu