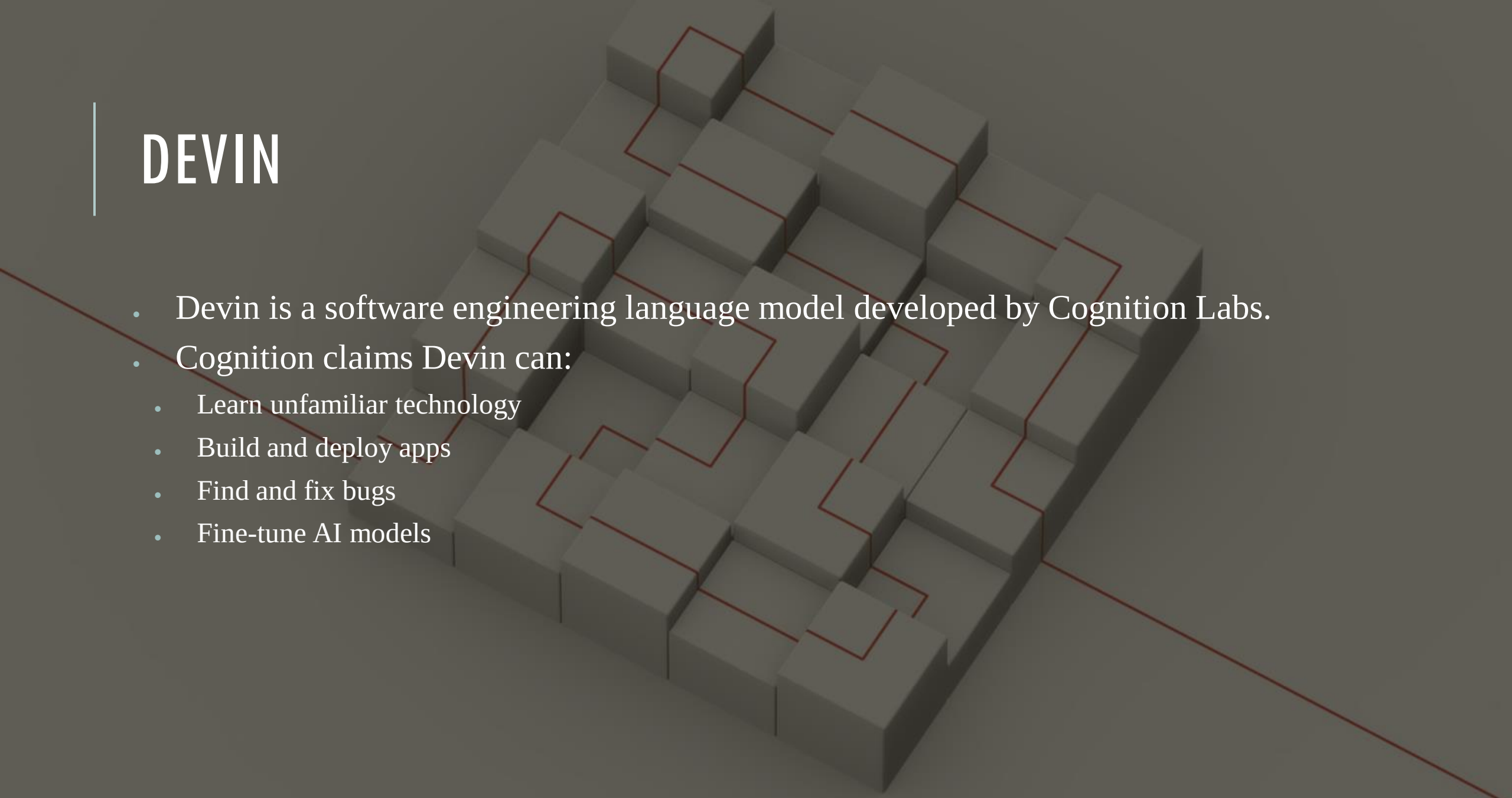
DEVIN: AI SOFTWARE ENGINEER

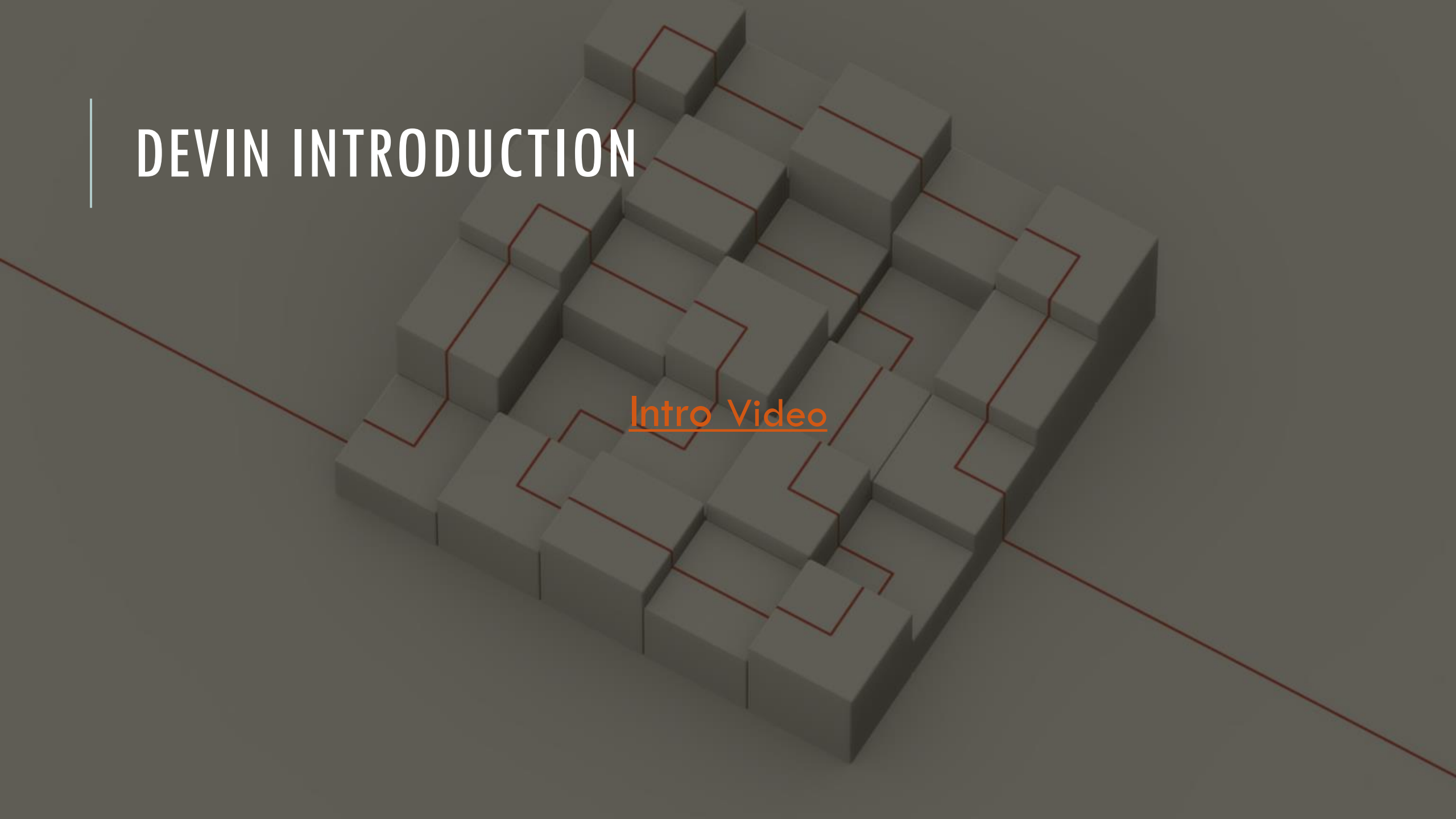
Dixon Ramey and Jadon Brutcher

DEVIN

- Devin is a software engineering language model developed by Cognition Labs.
- Cognition claims Devin can:
  - Learn unfamiliar technology
  - Build and deploy apps
  - Find and fix bugs
  - Fine-tune AI models

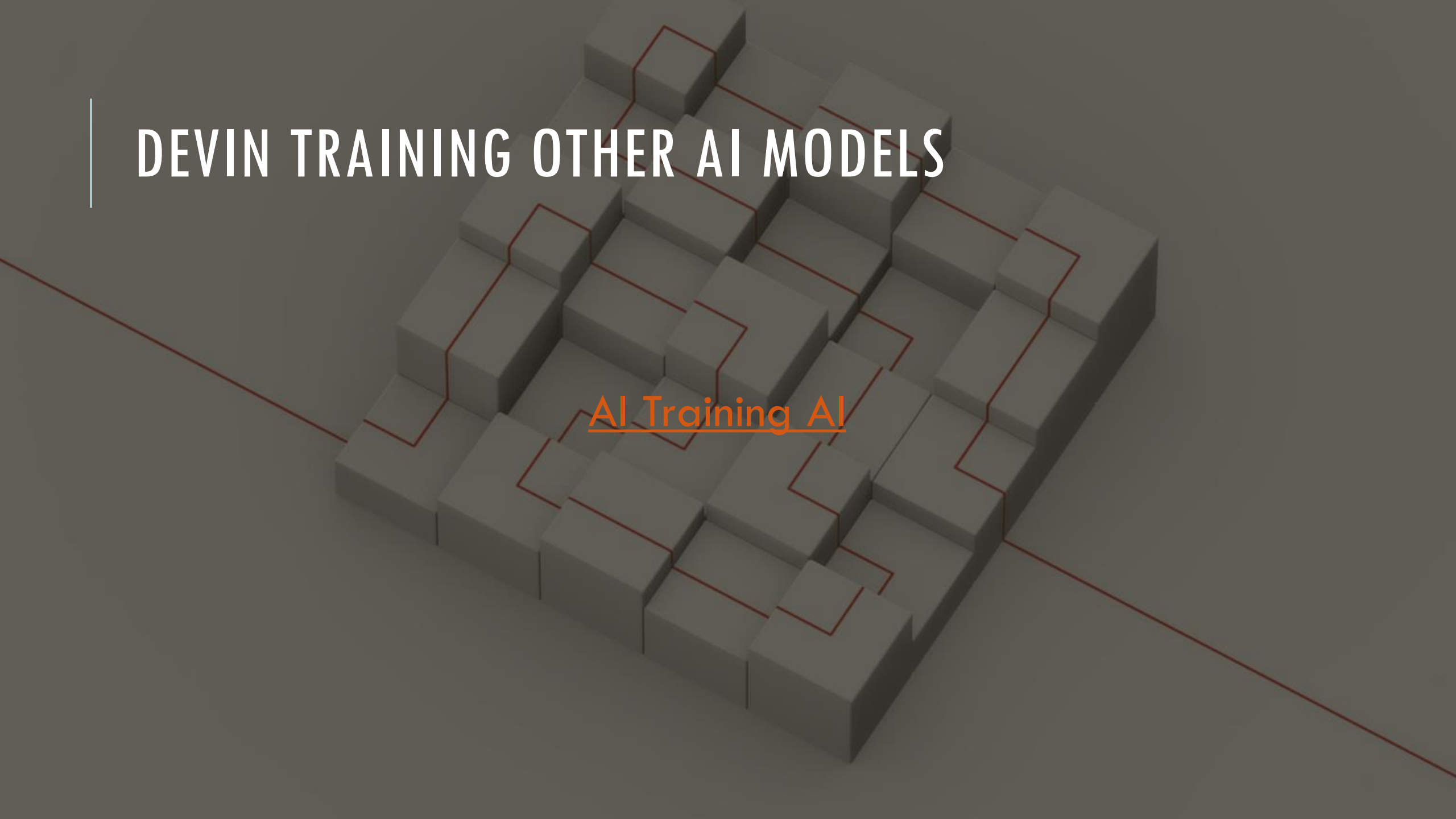


DEVIN INTRODUCTION



Intro Video

DEVIN TRAINING OTHER AI MODELS



AI Training AI

SWE BENCH

- SWE-bench consists of over 2,000 real world software problems drawn from GitHub issues and pull requests across 12 popular python repositories.
- Using GitHub issues and pull requests, the task then becomes to make a new pull request that solves the problem set out in the issue or original pull request.
- Evaluation is performed by unit test verification using post-PR behavior as the reference solution.

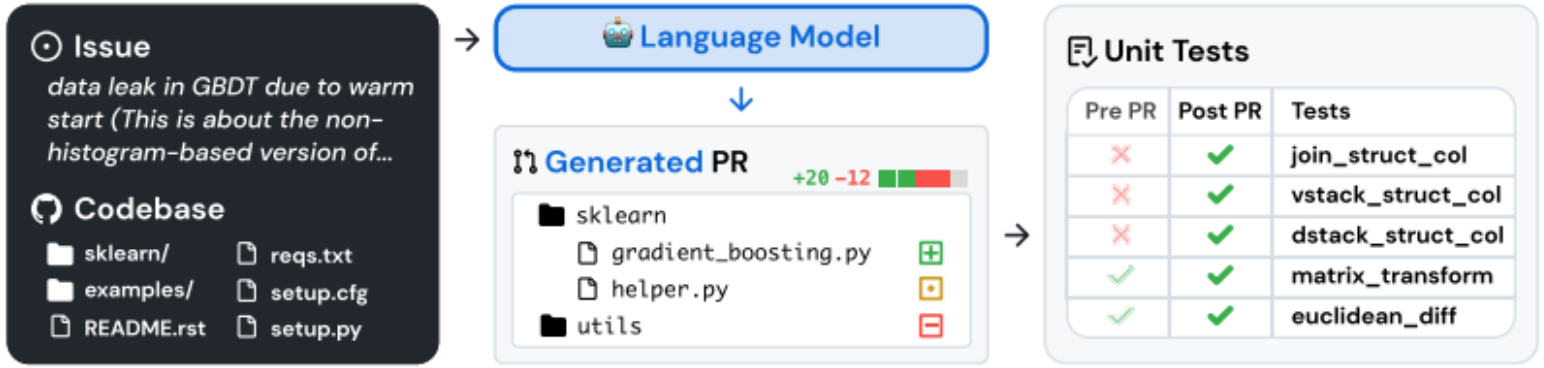


Figure 1: SWE-bench sources task instances from real-world Python repositories by connecting GitHub issues to merged pull request solutions that resolve related tests. Provided with the issue text and a codebase snapshot, models generate a patch that is evaluated against real tests.

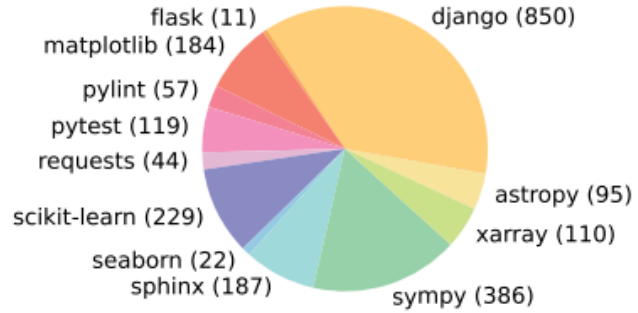


Figure 3: Distribution of SWE-bench tasks (in parenthesis) across 12 open source GitHub repositories that each contains the source code for a popular, widely downloaded PyPI package.

SWE BENCH LEADERBOARD

- Here is the current leaderboard as of this morning

Leaderboard

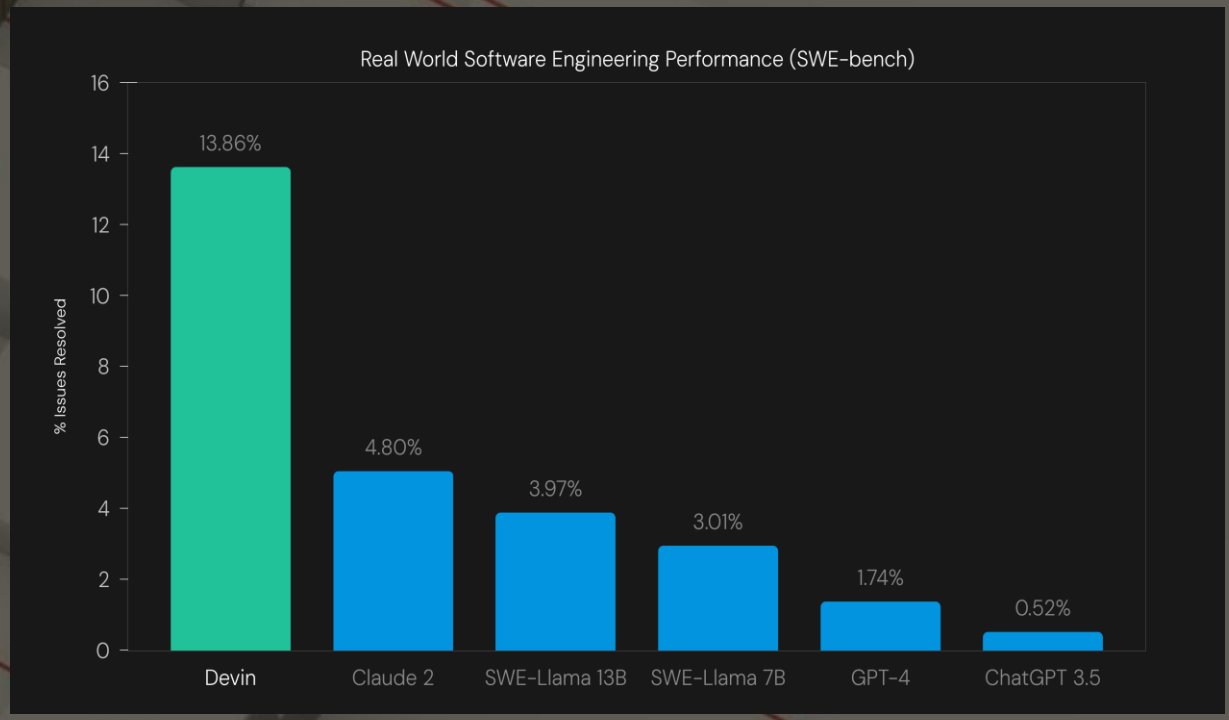
Model	% Resolved	Date
SWE-agent + GPT 4	12.29	2024-4-2
RAG + Claude 3 Opus	3.79	2024-4-2
RAG + Claude 2	1.96	2023-10-10
RAG + GPT 4	1.44	2024-4-2
RAG + SWE-Llama 13B	0.70	2023-10-10
RAG + SWE-Llama 7B	0.70	2023-10-10
RAG + ChatGPT 3.5	0.20	2023-10-10

The **% Resolved** metrics refers to the percentage of SWE-bench instances (2294 total) that were *resolved* by the model.

RESULTS & DEVIN GRAPH

Model	BM25 Retrieval		"Oracle" Retrieval	
	% Resolved	% Apply	% Resolved	% Apply
Claude 2	**1.96**	43.07	**4.80**	62.82
ChatGPT-3.5	0.17	26.33	0.52	21.80
GPT-4*	0.00	14.83	1.74	34.00
SWE-Llama 7b	0.70	51.74	3.01	65.52
SWE-Llama 13b	0.70	**53.62**	3.97	**66.78**

Table 18: We compare models against each other using the BM25 and oracle retrieval settings as described in Section 4. The main results table, Table 5, presents the results for the different models when using BM25 only. *Due to budget constraints we evaluate GPT-4 on a 25% random subset of SWE-bench in the "Oracle" and BM25 27K retriever settings only.

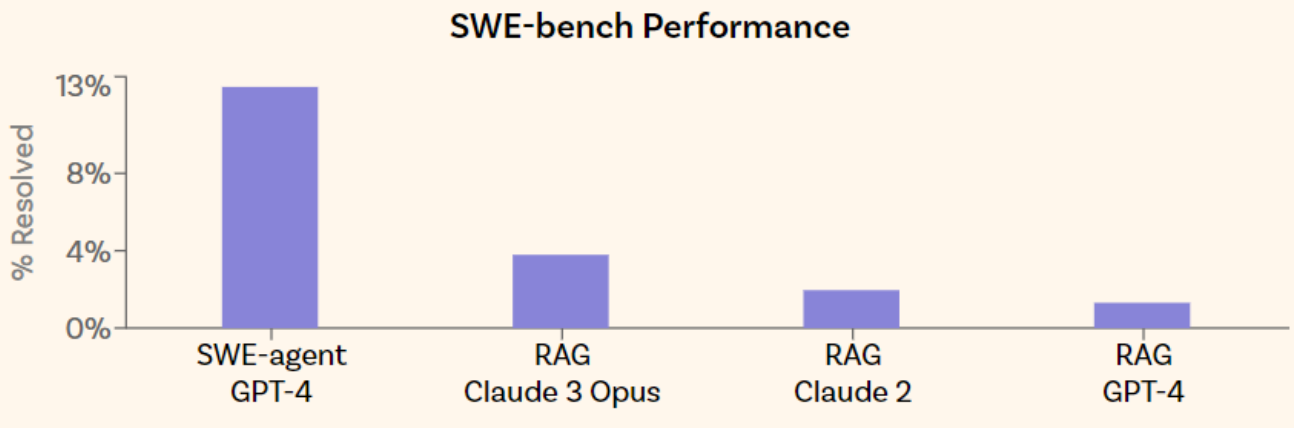


SWE-AGENT

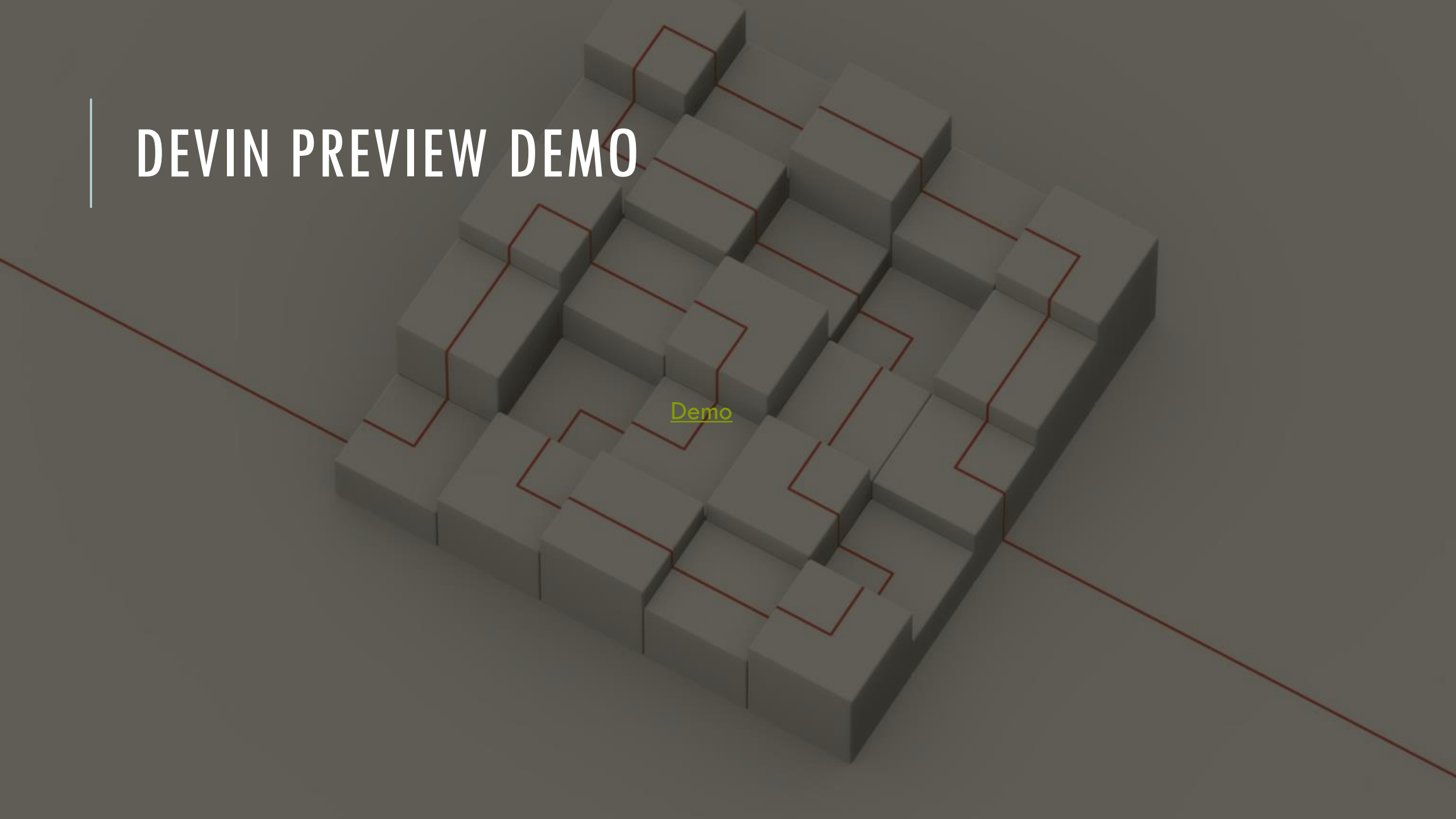
Additionally, developed by the same people at SWE-bench, SWE-agent is an interface that allows language models to fix bugs in GitHub repositories.

Utilizes full SWE-bench set and offers the code.

On the full SWE-bench test set, SWE-agent fixes 12.47% of issues, the new state-of-the-art result on the full test set.



DEVIN PREVIEW DEMO



Demo

IS DEVIN REAL?

- Currently access to Devin is not yet available, sparking claims that Devin is not real.
- Demo forces you to sign up for a waitlist
- No word from the company since announcement
- No recognition from SWE-bench
- Possibly suspicious website and videos

IMPACT

- If Devin is real
  - Could discourage new developers
  - New state of the art
  - Useful tool for developers
  - Increases competition for top AI firms
- If Devin isn’t real
  - Discourages new developers for no reason
  - Money from investors that could go to real work

POSSIBLE EXTENSIONS OF DEVIN

Currently there is speculation that Devin is not a real system or isn't all it claims to be.

A possible continuation from Cognition Labs would be evaluating Devin on the full SWE-bench, as well as providing access to the Devin demo so that users can experience it and provide feedback.

REFERENCES

Jimenez, C. E., Narasimhan, N., Pei, K., Press, P., Wetting, A., Yang, J., & Yao, S. (2024). SWE-BENCH: CAN LANGUAGE MODELS RESOLVE REAL-WORLD GITHUB ISSUES? https://arxiv.org/pdf/2310.06770.pdf

Jimenez, C. E., Narasimhan, N., Pei, K., Press, P., Wetting, A., Yang, J., & Yao, S. (2024). SWE-BENCH: CAN LANGUAGE MODELS RESOLVE REAL-WORLD GITHUB ISSUES? https://www.swebench.com/

Jimenez, C. E., Lieret, K., Narasimhan, N., Press, P., Wetting, A., Yang, J., & Yao, S. (2024). SWE-AGENT: Agent Computer Interfaces Enable Software Engineering Language Models. https://swe-agent.com/

Wu, S. (2024). Introducing Devin, the first AI software engineer. https://www.cognition-labs.com/

QUESTIONS???