

[Skip to main content](#)> [cs](#) > arXiv:2107.03374

quick links

- [Login](#)
- [Help Pages](#)
- [About](#)

Computer Science > Machine Learning

arXiv:2107.03374 (cs)

[Submitted on 7 Jul 2021 ([v1](#)), last revised 14 Jul 2021 (this version, v2)]

Evaluating Large Language Models Trained on Code

[Mark Chen](#), [Jerry Tworek](#), [Heewoo Jun](#), [Qiming Yuan](#), [Henrique Ponde de Oliveira Pinto](#), [Jared Kaplan](#), [Harri Edwards](#), [Yuri Burda](#), [Nicholas Joseph](#), [Greg Brockman](#), [Alex Ray](#), [Raul Puri](#), [Gretchen Krueger](#), [Michael Petrov](#), [Heidy Khlaaf](#), [Girish Sastry](#), [Pamela Mishkin](#), [Brooke Chan](#), [Scott Gray](#), [Nick Ryder](#), [Mikhail Pavlov](#), [Alethea Power](#), [Lukasz Kaiser](#), [Mohammad Bavarian](#), [Clemens Winter](#), [Philippe Tillet](#), [Felipe Petroski Such](#), [Dave Cummings](#), [Matthias Plappert](#), [Fotios Chantzis](#), [Elizabeth Barnes](#), [Ariel Herbert-Voss](#), [William Hebgen Guss](#), [Alex Nichol](#), [Alex Paino](#), [Nikolas Tezak](#), [Jie Tang](#), [Igor Babuschkin](#), [Suchir Balaji](#), [Shantanu Jain](#), [William Saunders](#), [Christopher Hesse](#), [Andrew N. Carr](#), [Jan Leike](#), [Josh Achiam](#), [Vedant Misra](#), [Evan Morikawa](#), [Alec Radford](#), [Matthew Knight](#), [Miles Brundage](#), [Mira Murati](#), [Katie Mayer](#), [Peter Welinder](#), [Bob McGrew](#), [Dario Amodei](#), [Sam McCandlish](#), [Ilya Sutskever](#), [Wojciech Zaremba](#)

[Download PDF](#)


We introduce Codex, a GPT language model fine-tuned on publicly available code from GitHub, and study its Python code-writing capabilities. A distinct production version of Codex powers GitHub Copilot. On HumanEval, a new evaluation set we release to measure functional correctness for synthesizing programs from docstrings, our model solves 28.8% of the problems, while GPT-3 solves 0% and

GPT-J solves 11.4%. Furthermore, we find that repeated sampling from the model is a surprisingly effective strategy for producing working solutions to difficult prompts. Using this method, we solve 70.2% of our problems with 100 samples per problem. Careful investigation of our model reveals its limitations, including difficulty with docstrings describing long chains of operations and with binding operations to variables. Finally, we discuss the potential broader impacts of deploying powerful code generation technologies, covering safety, security, and economics.

Comments: corrected typos, added references, added authors, added acknowledgements

Subjects: **Machine Learning (cs.LG)**

Cite as: [arXiv:2107.03374](https://arxiv.org/abs/2107.03374) [cs.LG]
(or [arXiv:2107.03374v2](https://arxiv.org/abs/2107.03374v2) [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2107.03374>

 Focus to learn more

arXiv-issued DOI via DataCite

Submission history

From: Mark Chen [[view email](#)]

[v1] Wed, 7 Jul 2021 17:41:24 UTC (1,466 KB)

[v2] Wed, 14 Jul 2021 17:16:02 UTC (1,467 KB)

☐ Bibliographic Tools

Bibliographic and Citation Tools

☐ Bibliographic Explorer Toggle

Bibliographic Explorer ([What is the Explorer?](#))

☐ Litmaps Toggle

Litmaps ([What is Litmaps?](#))

☐ scite.ai Toggle

scite Smart Citations ([What are Smart Citations?](#))

☒ Code, Data, Media

Code, Data and Media Associated with this Article

☐ Links to Code Toggle

CatalyzeX Code Finder for Papers ([What is CatalyzeX?](#))

☐ DagsHub Toggle

DagsHub ([What is DagsHub?](#))

☐ Links to Code Toggle

Papers with Code ([What is Papers with Code?](#))

☐ ScienceCast Toggle

ScienceCast ([What is ScienceCast?](#))

☐ Demos

Demos

☐ Replicate Toggle

Replicate ([What is Replicate?](#))

☐ Spaces Toggle

Hugging Face Spaces ([What is Spaces?](#))

☐ Related Papers

Recommenders and Search Tools

☐ Link to Influence Flower

Influence Flower ([What are Influence Flowers?](#))

☐ Connected Papers Toggle

Connected Papers ([What is Connected Papers?](#))

☐ Core recommender toggle

CORE Recommender ([What is CORE?](#))

☐ IArxiv recommender toggle

IArxiv Recommender ([What is IArxiv?](#))

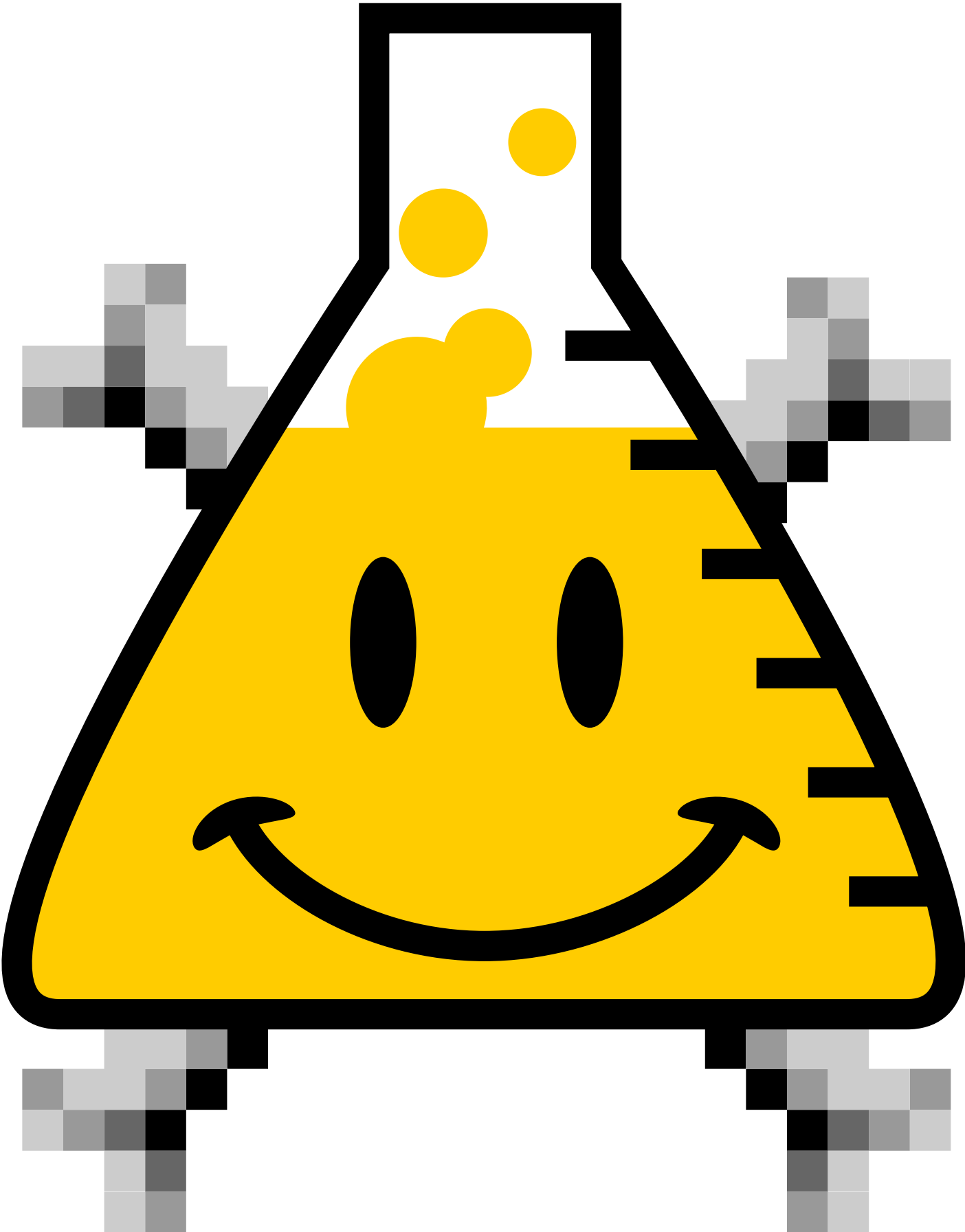
☐ About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs](#).



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))