**1. Benchmarks and Metrics for Evaluations of Code Generation: A Critical Review**

<https://arxiv.org/pdf/2406.12655v1>

**2. L2CEval: Evaluating Language-to-Code Generation Capabilities of Large Language Models**

<https://ui.adsabs.harvard.edu/abs/2023arXiv230917446N/abstract>

**3. Fully Autonomous Programming with Large Language Models**

<https://dl.acm.org/doi/10.1145/3583131.3590481>

**4. "** **SPoC: Search-based Pseudocode to Code "**

<https://arxiv.org/abs/1906.04908>

THIS PAPER IS CITED IN PRAXIS. SEE HOW – AND IF THE SUMMARY CAN BE USEFUL

**5. "Evaluating Large Language Models Trained on Code"**

* **Authors:** Mark Chen, Jerry Tworek, Heewoo Jun, et al.
* **Published by:** arXiv, 2021
* **Overview:** This paper presents the evaluation of Codex, a large language model trained on code. The authors use the HumanEval dataset, which includes unit tests, to assess the functional correctness of the code generated by Codex. The paper provides detailed results on the model's performance across various programming tasks and discusses the role of unit tests in validating code generation.
* **Link:** [arXiv:2107.03374](https://arxiv.org/abs/2107.03374)