**Summary of our Work**

**Yakov Schory**,**Lia Roiyhvarg** ResearchGenAI

**David Aslanyan, Michael Kulikov** CodeCrusaders

The paper "Evaluating Large Language Models Trained on Code" introduced Codex (a GPT-based language model fine-tuned on publicly available code). The study evaluated Codex’s ability to generate functional code from docstrings and compared its performance with previous models like GPT-3 and GPT-J on the HumanEval benchmark — a new dataset created to assess the correctness of code by using unit tests. Evaluations were performed using the pass@k metric that uses unit tests to evaluate code correctness.

The second paper, "Capturing Failures of Large Language Models via Human Cognitive Biases," delves into the challenge of mitigating biases in LLMs. It examines various cognitive biases, such as the framing effect, anchoring effect, and availability heuristic, and evaluates how models like GPT-3 and Codex are influenced by these biases. The paper demonstrates how these biases manifest differently in AI models, highlighting the importance of addressing them to improve the accuracy and fairness of LLM-generated outputs.

Incorporating insights from these papers, our solution integrates key findings into a practical framework. We utilized the pass@k metric to compare the performance of the GPT-2 model with a fine-tuned SFT (supervised fine tuning) model. Our objective was to generate optimized outputs with reduced cognitive biases. To achieve this, we employed the HumanEval dataset as our primary source for programming problems. By leveraging natural language processing algorithms, the tool aims to enhance the accuracy and functionality of AI-generated code while addressing the biases identified in the second paper and improving upon the code generation techniques discussed in the first paper. This approach not only refines prompts but also evaluates outputs against a benchmark dataset, ensuring that the final product benefits from both improved performance and reduced cognitive biases.