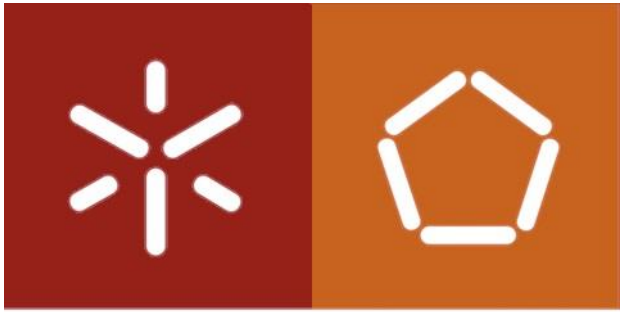


Towards a Platform for Benchmarking Large Language Models

Universidade do Minho
Escola de Engenharia

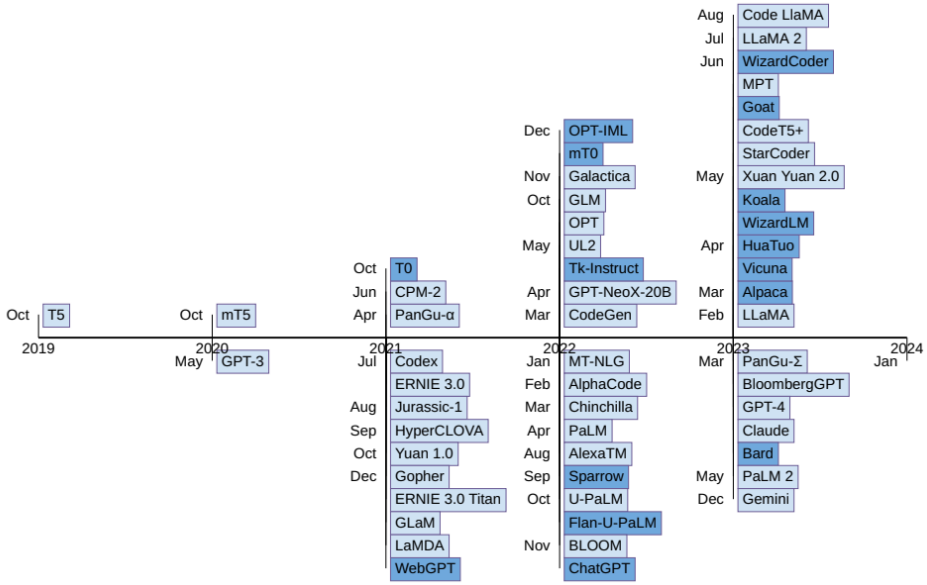
Simão Pedro Sá Cunha

Trabalho efetuado sob a orientação de
João Alexandre Baptista Vieira Saraiva
Francisco José Torres Ribeiro

Dissertação de mestrado em Engenharia Informática

Contextualização e motivação

- LLMs estão a mudar a forma como criamos *software.*
- Cada vez há mais modelos e todos eles consomem imensa energia.
- **Objetivo:** saber os seus consumos energéticos e identificar qual é o modelo mais eficiente.
- **De que forma?** Utilizar modelos como se se tratasse de uma linguagem de programação e *prompts (*instrução dada ao modelo*)* como um programa. Já existem *benchmarks (*conjuntos de *prompts)* para analisar a eficácia dos modelos e serão usados no estudo.



Retirado de *Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2023*.

Roadmap previsto

- Revisão literária de LLMs, *benchmarks* e *frameworks* de medição de consumo de energia;
- Identificação de recursos computacionais;
- Codificar *scripts;*
- Recolha de dados;
- Análise multi-critério que avalia o consumo de energia e de memória, tempo de execução e precisão dos modelos*;*
- Criação de um ambiente para fácil recreação do estudo.

Problemas e desafios identificados

Problemas

Armazenamento escasso para guardar os modelos

Obter RAM e GPU melhores para recolha de dados mais rápida

Processos ativos que resultam em consumo de energia desnecessário

Solução

Acesso a servidor por SSH com mais armazenamento, com melhor RAM e GPU e com o mínimo de processos ativos

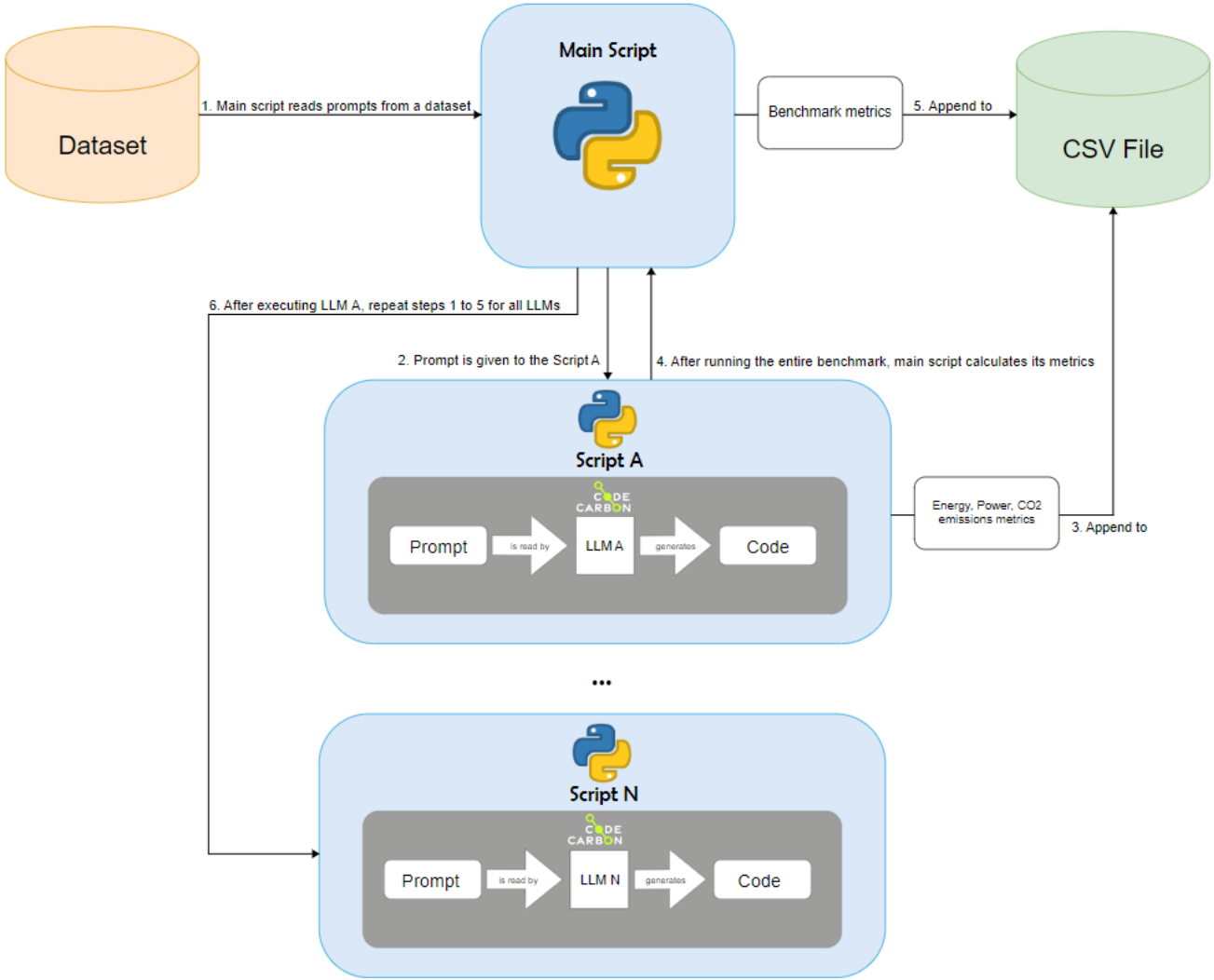
Metodologia

Execução dos modelos localmente – existe uma biblioteca Python que facilita o processo

Medição da eficácia dos modelos – contém problemas de programação em C++, Go, Python, Java e JavaScript

Ferramenta de medição do consumo de energia

Arquitetura do sistema

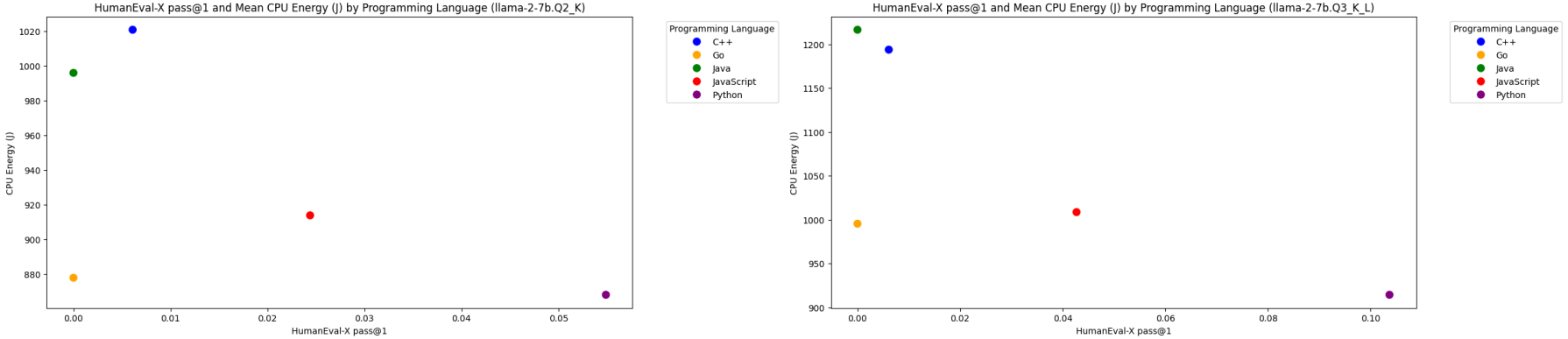


Resultados preliminares

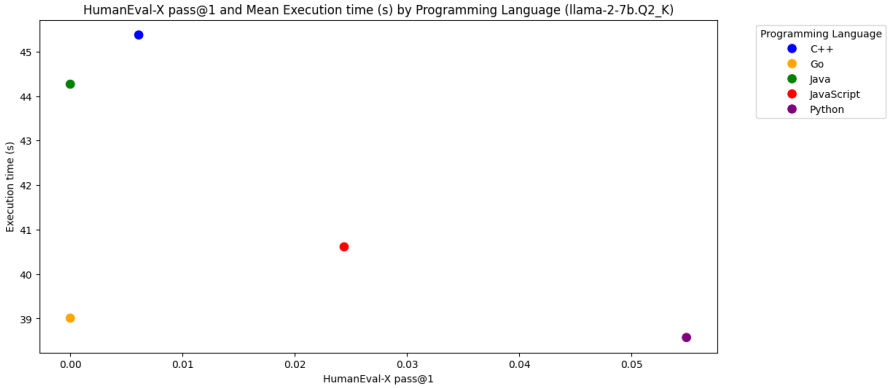
Operating System	Linux Ubuntu 22.04
Processor	Intel(R) Core(TM) i7-8750H
Clockspeed	2.2 GHz
Turbo Speed	4.1 GHz
Cores	6
Threads	12
RAM	16GB
RAM Speed	2666 mt/s
Cache Size	L1: 384K, L2: 1.5MB, L3: 9MB

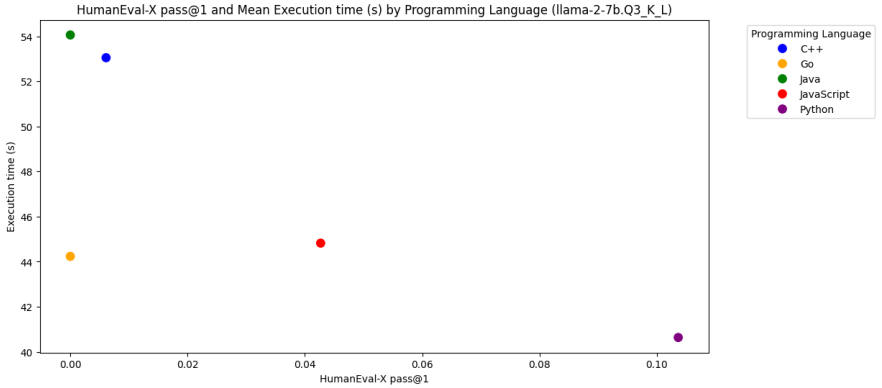
Especificações da máquina utilizada

Resultados preliminares – Energia consumida



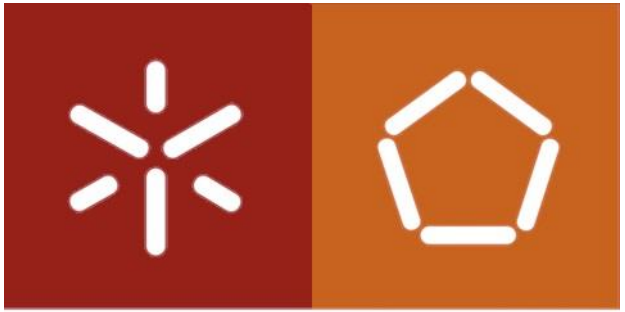
Resultados preliminares – Tempo de execução





Conclusão e trabalho futuro

- Nesta fase obtivemos alguns resultados preliminares, mas pouco fiáveis;
- A solução passa na obtenção uma máquina com acesso SSH com melhor *hardware* que o utilizado;
- Próximos passos irão focar-se na adição de mais *benchmarks* como o CyberSecEval (sobre cibersegurança) e o MBPP (sobre problemas de programação apenas em Python);
- Também será criado um ambiente – e.g. numa imagem Docker – para facilitar a reprodução dos resultados obtidos.



Towards a Platform for Benchmarking Large Language Models

Universidade do Minho
Escola de Engenharia

Simão Pedro Sá Cunha

Trabalho efetuado sob a orientação de
João Alexandre Baptista Vieira Saraiva
Francisco José Torres Ribeiro

Dissertação de mestrado em Engenharia Informática