Universidade do Minho
Escola de Engenharia

Plano de Trabalho de Dissertação

Ano Letivo 2023/2024

Nome Estudante	Simão Pedro Sá Cunha
N.º Estudante	A93262
Curso	Mestrado Integrado em Engenharia Informática
Título da Dissertação (em Português)	Avaliação comparativa de modelos de linguagem grandes
Título da Dissertação (em Inglês)	Towards a Platform for benchmarking Large Language Models

Enquadramento e Motivação (150 - 200 palavras)

Large Language Models (LLM), tais como o GPT-4 que suporta o ChatGPT, estão a mudar a forma como desenvolvemos software. Estes modelos são capazes de apoiar os programadores no desenvolvimento de código, tal como fazem os IDE avançados que utilizam o apoio do coPilot (tal como o ChatGPT usa um LLM no seu core). Devido ao grande sucesso destes LLMs, mais modelos têm surgido (e.g. LLAMA-2 da Meta), variando nas suas dimensões e capacidades. Criar e treinar LLMs é uma tarefa que exige muita computação, obrigando a um grande consumo de energia. De facto, os princípios de sustentabilidade em engenharia de software e LLM parecem caminhar em direções opostas.

Uma vez que vão existindo mais LLMs é importante analisar as suas performances, não apenas gerar/completar um programa numa qualquer linguagem de programação, mas também a sua performance energética. Assim pretende-se definir um conjunto de problemas onde LLMs são habitualmente usados (expressos como prompts que são os seus inputs). Nesta tese será feito depois um estudo sobre a performance de todos os LLMs disponíveis para resolver esses problemas. A performance energética de cada LLM será ainda medida, de modo a sabermos qual o modelo mais sustentável a resolver um determinado problema.

Objetivos e Resultados Esperados (150 - 200 palavras)

O principal objetivo desta dissertação consiste numa análise comparativa de LLMs tendo em conta o seu consumo energético, tempo de execução e acurácia na resposta devolvida de acordo com *benchmarks* já conhecidos. Para tal, será necessário cumprir os seguintes tópicos:

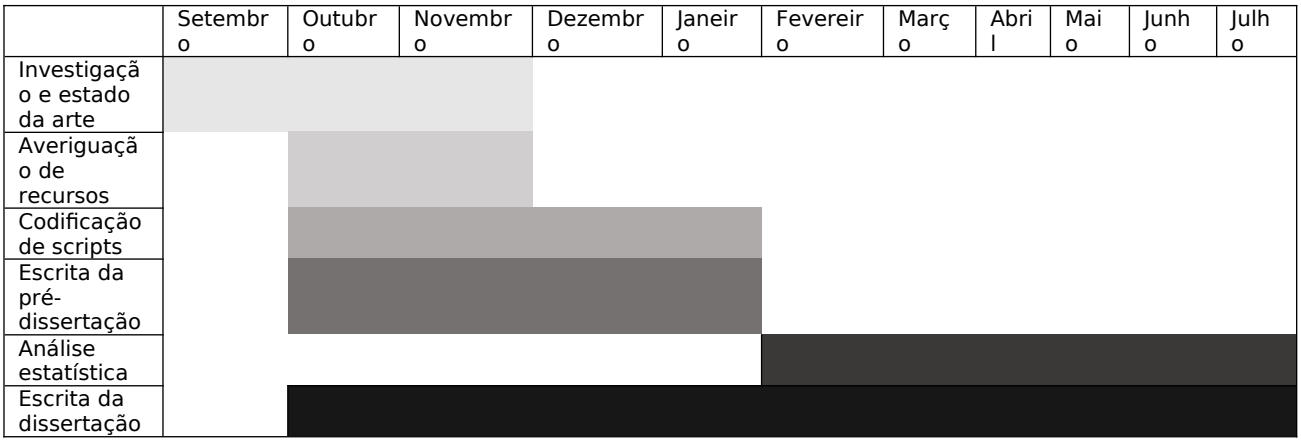
- Recolha de bibliografia sobre LLMs, benchmarks aplicados a LLMs e de *frameworks* sobre medição de consumo de energia da máquina que estiver a executar os LLMs;
- Averiguação dos recursos computacionais necessários para efetuar a

experiência;
- Codificação de *scripts* que apliquem os vários LLMs de forma automática de modo a que seja acessível a recriação dos resultados obtidos;
- Obtenção de um ficheiro (e.g. CSV - *Comma Separated Values*) com todas as medições efetuadas;
- Análise estatística dos resultados obtidos.

Os principais resultados que se esperam desta dissertação são, para cada benchmark considerado:

a. Identificar o LLM com o melhor tempo de execução e com o menor consumo de energia, excluíndo a acurácia obtida;
b. Identificar o LLM com o menor consumo de energia e com a maior acurácia obtida, excluíndo o tempo de execução;
c. Identificar o LLM com a maior acurácia obtida e com o menor tempo de execução e menor consumo de energia possíveis.

Calendarização



Referências Bibliográficas (5 - 10 referências)

[1] Baptiste R. *et al* (2023) “Code Llama: Open Foundation Models for Code” - https://ai.meta.com/research/publications/code-llama-open-foundation-models-for-code/ (consultado em outubro 2023)
[2] Chen M. *et al* (2021) “Evaluating Large Language Models Trained on Code” - https://arxiv.org/abs/2107.03374 (consultado em outubro 2023)
[3] Touvron H. *et al* (2023) “Llama 2: Open Foundation and Fine-Tuned Chat Models” - https://arxiv.org/abs/2307.09288 (consultado em outubro 2023)
[4] Hendrycks D. *et al* (2020) “Measuring Massive Multitask Language Understanding” - https://arxiv.org/abs/2009.03300 (consultado em outubro 2023)
[5] Marcus H. et al (2012) "Measuring energy consumption for short code paths using RAPL". SIGMETRICS Perform. Eval. Rev., 40(3):13–17 - https://dl.acm.org/doi/10.1145/2425248.2425252 (consultado em outubro 2023)

Justificação de Coorientação (se aplicável)

Assinaturas

Estudan

Assinado por: **Simão Pedro Sá Cunha**
Num. de Identificação: 15179799
Data: 2023.10.15 17:38:07+01'00'

Orientador (tal como previsto no ponto 1 do Artigo

Assinado por: **João Alexandre Baptista Vieira Saraiva**
Num. de Identificação: 07784474
Data: 2023.10.12 16:42:33 +0100

Diretor do Ciclo de

Orientador (tal como previsto no ponto 3 do Artigo 169.º do RAUM. Neste caso, é obrigatório existir um Orientador pelo ponto 1 do Artigo 169.º do RAUM)

Assinatura digital qualificada com Cartão de Cidadão ou Chave Móvel Digital. Para os estudantes, nos casos em que tal não seja possível, os mesmos deverão imprimir este plano, assinar manualmente e, após digitalização, os restantes intervenientes usam a assinatura digital qualificada.

O plano de trabalho deve ser preenchido *offline* e realizado o *upload* do mesmo, depois de assinado, no formulário do requerimento de pedido de admissão à dissertação, disponível em http://dissertacao.eng.uminho.pt