Software Quality Engineering, 2025-2026
Homework #3: Software performance benchmarking

DEADLINE

- One week before the date of the group oral exam.

ASSIGNMENT

You are required to benchmark the inference time of one or more LLMs that are instructed for question-answering (chatbot). The main goals are: (i) benchmark the inference time of selected LLMs executed locally, (ii) collect the related latency metrics, and (iii) analyze the metrics to characterize the performance behavior of the models.

Required steps:

- Choose one or more language models that can safely run on your local machine. Models may be selected from the Ollama catalogue or the Hugging Face model repository.
- Define the benchmarking methodology (number of runs and workload) and the execution environment (e.g., hardware used, and software stack).
- Write the benchmarking scripts to run the experiments and collect the LLM-related latency metrics (included but not limited to TTFT, ITL, E2E latency).
- Analyse the metrics collected. Your analysis has to explore questions as:
    - Is there a relationship between metrics and input prompt?
    - Is there a relationship between metrics and output length?
    - Do metrics change when using different models/prompts?
    - Do results vary when executing multiple times with the same workload?

DELIVERABLES

- A detailed document explaining all steps performed during the homework (as listed above), including methodology, results, and analysis.
- GitHub Repository containing:
    - source code for running the benchmarking experiments;
    - source code for analysing the collected performance metrics.
- The performance measurements collected during the benchmarking (json/csv formats).

PROVIDED MATERIAL

We will provide each group with a dataset of prompts (LLM inputs) that you are expected to use for the benchmarking experiments.

USEFUL RESOURCES

- Ollama models catalogue
- LLM inference latency metrics