UNIVERSIDADE DE AVEIRO
DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA

Information and Coding (2025/26)

Lab work nº 3 — Due: 19 Dec 2025

- In this work, you will have to find the best way to compress the file `https://huggingface.co/Qwen/Qwen2-0.5B/resolve/main/model.safetensors`. This file contains the model parameters of a large language model (LLM) and is almost 1 GB long. You can use all means that you find appropriate to compress it, including, of course, a deep analysis of the file contents and structure. Your grade will be related to the combination of compression ratio, computation time to compress / decompress the file, and memory usage.

- Elaborate a concise report, where you describe all the relevant steps and decisions taken to arrive at your solution. Include measures of processing time and compression ratio, for several operation points (more compression but also more computing time versus less compression but faster). Do not forget to test existing compressors and to report the results that you obtain with them (compression ratio, time, memory usage).

- Prepare a presentation to show your work. It should be at most of 10 minutes and will be presented during the classes of 22nd December.