Section 2.

**USE OF LOCAL LANGUAGE MODELS**

Popov R. O.

Oles Honchar Dnipro National University

Scientific adviser Karpenko N. V.

**Relevance**. The impact of large language models (LLMs) is obvious: for users, it is a tool for learning and analysis, and researchers are constantly developing new architectures that show better results. LLMs have also become important for business, as many applications use APIs from OpenAI, Anthropic, Mistral, and other companies. This is convenient and efficient, as models require large resources to run.

However, using remote models via APIs has disadvantages, such as privacy issues, limited customization, and others. The question arises: is it possible to run LLM locally and what advantages will it bring compared to remote models? How will it affect the efficiency of AI applications?

**Objective**. To consider tools for local launch of language models and analyze their capabilities in comparison with remote models of IT companies.

**Main points**. Running large language models requires expensive equipment and high energy consumption. Instead, the topic of small language models (SLMs), which require less memory and have a high output generation speed, is gaining popularity [1].

Three main factors affect the amount of resources to run SLM: number of parameters, quantization, and architecture. The number of parameters is usually measured in billions, and companies release models with different numbers of parameters, such as Google Gemma 2 with 2 billion, 9 billion, and 27 billion parameters. Quantization can reduce the size of the model and speed up arithmetic operations, which is achieved by simplifying the representation of numbers (8, 4, or even 2 bits) [2].

According to the results of SLM testing on different data sets, it is the architecture and the date of model release that have the greatest impact on the quality of generation. Different architectures are adapted for different tasks, and the models of the current year are several times better than the models of the previous year [1]. Microsoft's internal experiment showed that SLMs are quite efficient and have ten times lower maintenance costs than LLMs [3].

The following models are currently popular SLMs: Microsoft Phi3.5, Alibaba Qwen2.5, Google Gemma 2, Meta Llama 3.2, and others. Usually, most SLMs and LLMs are stored on the Hugging Face service, and the de facto standard for running models locally is the llama.cpp program. Integration of local models into an existing AI application can be completely seamless, as llama.cpp (and other programs) can provide their services via an HTTP API compatible with the OpenAI API.

**Conclusions**. We have considered the issue of running language models locally. The advantage of language models that can be run locally is the complete preservation of data confidentiality, in contrast to using remote language models through the API.

There are quite a few SLMs for solving various tasks that are freely available, they can be downloaded from Hugging Face and run locally through llama.cpp.

The analysis shows that for information processing tasks, SLMs are more efficient than LLMs and require much less computing resources.

**List of references**

1. Lu Z., Li X., Cai D. et al. Small Language Models: Survey, Measurements, and Insights. 09.2024. DOI:10.48550/ARXIV.2409.15790.

2. Li S., Ning X., Wang L. et al. Evaluating Quantized Large Language Models. 28.02.2024. DOI:10.48550/ARXIV.2402.18158.

3. Li B., Zhang Y., Bubeck S. et al. Small Language Models for Application Interactions: A Case Study. 05.2024. DOI:10.48550/ARXIV.2405.20347.

**Information about the authors**

Popov Ruslan O., student of the Department of Electronic Computers, M.T. +380959241172, popov\_r@365.dnu.edu.ua

Karpenko Nadiia Valeriivna, Associate Professor of the Department of Electronic Computers, Oles Honchar Donetsk National University, PhD in Physics and Mathematics, Associate Professor, M.T. +380963155031, karpenko\_n@365.dnu.edu.ua