# Лабораториска вежба 5 – Збогатување на LLM со знаење и Retrieval Augmented Generation

## Задачи

Во оваа лабораториска вежба ќе се применуваат 3 податочни множества: *RAG-MINI-Wikipedia*, *CommonsenseQA* и *Yelp_parallel*. Множеството *RAG-Mini-Wikipedia* е наменето за одговарање прашања со кратки одговори кои може да се најдат во Wikipedia. Тоа е составено од прашања, одговори, и извадоци од Wikipedia кои служат за дополнителен контекст за одговарање на прашањата. *CommonSenseQA* е наменето за одговарање на прашања преку избор од понудени одговори (multiple choice question answering). Составено е од прашања и пет понудени одговори (еден точен и четири неточни). *Yelp_parallel* е наменето за трансформација на реченици кои содржат негативен сентимент во реченици кои содржат позитивен сентимент. Составено е од парови реченици кои содржат негативен и позитивен сентимент.

### Задача 1 (25 поени)

Користејќи го моделот LLaMA-2 со квантизација со 4bits и техниката RAG, генерирајте одговор за секое прашање од податочното множество *CommonSenseQA*. Од базата на знаење *GenericsKB* изберете контекст од вкупно 5 документи за секое прашање со семантичко пребарување и векторска репрезентација на документите со моделот *all-MiniLM-L6-v2*.

За евалуација користете ги метриките: *BLEU* и *BERTScore*.

Испробајте со различен број на документи како контекст. Како влијае бројот на документи врз резултатите?

Испробајте и друг модел за векторска репрезентација на документите (пр. DistilRoBERTa). Како влијае промената на моделот врз резултатите?

Испробајте и генерирање одговор на прашањата без користење контекст т.е. zero-shot prompting. На кој начин се добиваат подобри резултати?

### Задача 2 (25 поени)

Користејќи го моделот LLaMA-2 со квантизација со 4bits и техниката RAG, генерирајте одговор за секое прашање од податочното множество *RAG-Mini-Wikipedia*. Изберете контекст од вкупно 5 документи за секое прашање со семантичко пребарување и векторска репрезентација на документите со моделот *all-MiniLM-L6-v2*.

За евалуација користете ги метриките: *BLEU* и *BERTScore*.

Испробајте со различен број на документи како контекст. Како влијае бројот на документи врз резултатите?

Испробајте и друг модел за векторска репрезентација на документите (пр. DistilRoBERTa). Како влијае промената на моделот врз резултатите?

Испробајте и генерирање одговор на прашањата без користење контекст т.е. zero-shot prompting. На кој начин се добиваат подобри резултати?

## Задача 3 (25 поени)

Користејќи го моделот LLaMA-2 со квантизација со 4bits и техниката RAG, трансформирајте ги реченициите кои содржат негативен сентимент од податочното множество *Yelp_parallel* во реченици со позитивен сентимент. Изберете контекст од вкупно 5 документи за секое прашање со семантичко пребарување и векторска репрезентација на документите со моделот *all-MiniLM-L6-v2*.

За евалуација користете ги метриките: *BLEU* и *BERTScore*.

Испробајте со различен број на документи како контекст. Како влијае бројот на документи врз резултатите?

Испробајте и друг модел за векторска репрезентација на документите (пр. DistilRoBERTa). Како влијае промената на моделот врз резултатите?

Дали овој пристап е подобар споредено со претходните лабораториски вежби?