

**МОЛБА
ЗА ОДОБРАВАЊЕ ТЕМЕ МАСТЕР РАДА**

Молим да ми се одобри израда мастер рада под насловом:

Методе за аутоматску оптимизацију и структурирање упита за велике језичке моделе

Значај теме и области:

Популарност великих језичких модела (енг. *Large Language Models*) у оквиру области вештачке интелигенције поставила је инжењеринг упита (енг. *Prompt Engineering*) као битан фактор за постизање високе рачунарске ефикасности и робусности у применама обраде природног језика. Значај ове теме огледа се у потреби за дубљим разумевањем начина на који синтаксне и семантичке варијације у упиту утичу на стабилност, (не)детерминистичко понашање и тачност одговора које дају велики језички модели. Мастер рад би се бавио аутоматском оптимизацијом и прилагођавањем корисничких упита, који у пракси често садрже висок ниво семантичког шума, редундантност, лош редослед инструкција или логичке контрадикторности. Значај саме теме лежи у потенцијалу да се кроз чишћење, преслагање и друго прилагођавање улазног упита смањи потрошња рачунарских ресурса и истовремено подигне квалитет добијеног одговора.

Специфични циљ рада:

У раду ће се анализирати и практично тестирати неколико комплементарних приступа. Они обухватају детерминистичке лингвистичке хеуристике за примарно прочишћавање упита и уклањање шума, затим методе засноване на пресликавању реченица у векторски простор (употребом семантичких ембединга) ради елиминисања редундансе, као и потенцијално структурално моделовање помоћу графовских репрезентација у циљу дефинисања оптималног логичког редоследа инструкција. Крајња архитектура система биће обликована кроз појединачну примену или спој оних техника које се у експерименталној фази покажу као најефикасније за побољшање квалитета и тачности добијених одговора. Циљ рада био би да се на репрезентативном јавно доступном скупу података (као што је *BPO – Black-Box Prompt Optimization*), прецизно квантификује степен побољшања квалитета добијених одговора након оптимизације упита. Експерименти би се спровели над актуелним језичким моделима отвореног кода. Како би се контролисао недетерминизам генеративних модела, хиперпараметар температуре биће фиксиран на вредност 0. Квалитет излазних одговора мерио би се у односу на референтне одговоре употребом стандардних НЛП метрика (као што су *BLEU*, *ROUGE* и *BERTScore*), као и кроз семантичку евалуацију помоћу језичких модела у улози судије.

Остале битне информације:

Сав развијени програмски код и ресурси ће бити постављени на јавно доступни *GitHub* репозиторијум.

Драгана Здравковић, 1020/23, Информатика
(име и презиме студ., бр. инд., ознака програма и модула)

(својеручни потпис студента)

(датум подношења молбе)

Сагласан ментор др Александар Картељ

(својеручни потпис ментора)

Чланови комисије

1. др Младен Николић
2. др Сташа Вујичић Станковић

Катедра рачунарство и информатику је сагласна са предложеном темом.

(шеф катедре)

(датум одобравања молбе)