

КОМИСИЈИ ЗА СТУДИЈЕ II СТЕПЕНА ЕЛЕКТРОТЕХНИЧКОГ ФАКУЛТЕТА У БЕОГРАДУ

Комисија за студије II степена, Електротехничког факултета у Београду, на својој седници одржаној 21.07.2026 године именовала нас је у Комисију за преглед и оцену мастер рада дипл. инж. Кристина Петрић под насловом „Анализа перформанси инференције отворених језичких модела кроз поређење хардвера, извршних окружења и инфраструктуре". Након прегледа материјала Комисија подноси следећи

ИЗВЕШТАЈ

1. Биографски подаци кандидата

Кристина Петрић је рођена 5.3.2001. године у Београду. Основну школу „Филип Кљајић" завршила је као ђак генерације а паралелно је похађала и завршила основну музичку школу „Ватрослав Лисински“, смер клавир. Уписала је Тринаесту гимназију у Београду коју је завршила као вуковац. Електротехнички факултет Универзитета у Београду уписала је 2020. године, на модулу Сигнали и системи. Током основних студија била је ангажована као демонстратор. Основне студије завршила је 2024. године са просечном оценом 9,48, одбравивши дипломски рад под насловом „Препознавање говора из ограниченог речника применом неуралних мрежа и LPC анализе" са оценом 10. Мастер академске студије на Електротехничком факултету у Београду уписала је на модулу Софтверско инжењерство а испите је положила просечном оценом 9,80. Тренутно је запослена у компанији Microsoft на позицији Software engineer након завршене стручне праксе у истој фирми.

2. Извештај о студијском истраживачком раду

Кандидаткиња Кристина Петрић је као припрему за израду мастер рада урадила истраживање релевантне литературе која се односи на перформансе извршавања великих језичких модела. Конкретно су анализирана постојећа истраживања и мерења перформанси инференције великих језичких модела на сродном хардверу, укључујући академску студију која на рачунару опремљеном процесором M2 Ultra и 192 GB обједињене меморије пореди алате MLX, MLC-LLM, Ollama, llama.cpp и PyTorch MPS, као и независна ван-академска мерења и збирне резултате брзине генерисања по моделу и чипу. Проучени су и теоријски темељи неопходни за тумачење резултата: архитектура трансформерских модела и механизам пажње, разлика између обједињене и класичне (дискретне) архитектуре меморије, фазе инференције (обрада улазног промпта и декодовање) и њихова различита рачунска и меморијска ограничењеност, KV кеш и груписана упитна пажња (GQA), кеширање префикса, квантизација тежина модела и груписана обрада захтева (batching). Истраживањем области је утврђено да велика већина постојеће литературе фокус ставља на процес обучавања модела или на оптимизацију инференције у серверском окружењу са практично неограниченим ресурсима, док знатно мање пажње посвећено остаје питању максималног искоришћења расположивог хардвера и система за извршавање при локалној, ивичној инференцији на потрошачком уређају. У том простору је идентификована мотивација и научни допринос овог рада.

3. Опис мастер рада

Мастер рад обухвата 60 страна, са укупно 26 референци, 16 слика и 33 табеле. Рад садржи увод, 4 поглавља и закључак, списак коришћене литературе, слика и табела.

Прво поглавље представља увод у коме су описани предмет, циљ и мотивација рада.

Друго поглавље даје преглед постојећих истраживања и мерења перформанси инференције великих језичких модела на сродном хардверу, чиме се овај рад позиционира у односу на већ објављене резултате, укључујући и критички осврт на методолошке разлике између ове студије и независно објављених мерења. Треће поглавље уводи основне појмове из области отворених језичких модела и породице модела Llama 3.2 коришћене у раду, образлаже избор извршних окружења Ollama и llama.cpp сервера, објашњава разлику између обједињене и класичне архитектуре меморије, уводи основне концепте инференције (фазе инференције, KV кеш и груписану упитну пажњу, кеширање префикса, квантизацију и груписану обраду захтева) и метрике перформанси коришћене у раду, и на крају описује скупове података (ShareGPT, TriviaQA и скуп података корисничке подршке) коришћене за генерисање тестова.

Четврто поглавље описује архитектуру система коришћеног за мерење, укључујући хардверске платформе: рачунар са процесором Apple M1 Pro и виртуелну машину инфраструктуре Microsoft Azure.

Пето поглавље детаљно приказује и тумачи резултате свих седам спроведених контролисаних експеримената, поређаних према редоследу извршавања: поређење централног и графичког процесора, поређење система за извршавање Ollama и llama.cpp сервера, утицај кеширања префикса, утицај квантизације модела, поређење величине модела, утицај оркестрације системом Kubernetes на платформи Apple M1 Pro и контролна провера истог питања на инфраструктури Microsoft Azure са стварним графичким процесором произвођача NVIDIA.

Шесто, закључно поглавље сумира најважније доприносе рада, критички сагледава добијене резултате и ограничења спроведене анализе и даје смернице за даља истраживања.

4. Анализа са кључним резултатима

Мастер рад дипл. инж. Кристине Петрић бави се емпиријском анализом перформанси локалне инференције отворених језичких модела на хардверу архитектуре обједињене меморије, кроз седам независних, контролисаних експеримената спроведених на рачунару Apple M1 Pro уз завршну контролну проверу на инфраструктури Microsoft Azure са стварним графичким процесором произвођача NVIDIA. Сваки експеримент намерно је мењао само једну променљиву а кључна мерења су независно поновљена ради провере поузданости налаза.

Основни резултати рада су:

- 1) утврђивање да архитектура обједињене меморије знатно ублажава, али не уклања у потпуности, предност графичког процесора над централним
- 2) поређење система за извршавање Ollama и llama.cpp сервера, којим је показано да Ollama не остварује паралелизам под конкурентним оптерећењем за разлику од llama.cpp сервера
- 3) квантификација ефекта кеширања префикса и идентификација фактора од којих зависи стварна стопа поготка кеша у пракси
- 4) потврда да агресивна квантизација модела не доводи до мерљивог губитка тачности, уз знатно бољу брзину и мање заузеће меморије
- 5) поређење модела различите величине, које показује да мањи модел остварује знатно већу брзину уз тачност изједначену са већим моделом;
- 6) утврђивање да оркестрација системом Kubernetes сама по себи уводи занемарљив трошак, и да је ограничен приступ графичком процесору из контејнера на платформи Apple Silicon последица специфичног слоја виртуелизације, а не саме технологије оркестрације (налаз потврђен и на инфраструктури Microsoft Azure).

5. Закључак и предлог

Кандидаткиња Кристина Петрић је у свом мастер раду спровела методолошки ригорозну емпиријску анализу перформанси локалне инференције отворених језичких модела, обухватајући утицај хардвера, система за извршавање, квантизације, величине модела и начина оркестрације. Очекује се да рад послужи као корисна литература инжењерима и истраживачима који се у својој пракси сусрећу са изазовима локалне инференције великих језичких модела на ивичном и потрошачком хардверу.

На основу изложеног, Комисија предлаже Комисији за студије II степена Електротехничког факултета у Београду да рад дипл. инж. Кристине Петрић прихвати као мастер рад и кандидаткињи одобри јавну усмену одбрану.

Београд, 25.08.2026. године

Чланови комисије:

др Марија Пунт Ванредни професор
сагласан, 25.08.2026.

др Јелица Цинцовић Доцент
сагласан, 25.08.2026.