



Универзитет Сингидунум
Департман за последипломске студије

Мастер академске студије
Пословна економија: Маркетинг и трговина

Мастер теза

**Примена генеративне вештачке интелигенције и
великих језичких модела у имплементацији
интелигентних маркетинг решења**

Ментор:
др Ангелина Његуш

Кандидат:
Дејан Вујић - 2023400068

Београд, 2024.

Сажетак

Мастер теза истражује примену генеративне вештачке интелигенције (*Artificial Intelligence, AI*) и великих језичких модела (*Large Language Models, LLM*) у развоју интелигентних маркетиншких решења у савременом пословању. Студијско-истарживачки део истражује како алати вођени вештачком интелигенцијом, посебно интелигентни четботови засновани на архитектури *Retrieval-Augmented Generation (RAG)*, могу побољшати корисничку услугу и оперативну ефикасност. Кроз детаљну анализу технологија вештачке интелигенције, укључујући *Transformers* и *GPT* моделе, теза показује потенцијал генеративне вештачке интелигенције да револуционише интеракције купаца пружањем персонализованих и контекстуално релевантних одговора. Студија такође испитује економске и стратешке импликације интеграције вештачке интелигенције у маркетингу, нудећи увид у скалабилност и ефикасност ових технологија у апликацијама у стварном свету. Налази сугеришу да вештачка интелигенција, када се правилно примени, може значајно да побољша задовољство корисника и пословне перформансе, позиционирајући компаније на челу иновација у дигиталном добу.

Кључне речи:

Вештачка интелигенција | Генеративна вештачка интелигенција | Велики језички модели | Маркетинг | Интелигентни четбот

Abstract

The master's thesis investigates the application of generative artificial intelligence and large language models in the development of intelligent marketing solutions in modern business. The research arm explores how AI-driven tools, specifically intelligent chatbots based on the Retrieval-Augmented Generation (RAG) architecture, can improve customer service and operational efficiency. Through a detailed analysis of AI technologies, including Transformers and GPT models, the thesis demonstrates the potential of generative AI to revolutionize customer interactions by providing personalized and contextually relevant responses. The study also examines the economic and strategic implications of integrating artificial intelligence into marketing, offering insights into the scalability and effectiveness of these technologies in real-world applications. The findings suggest that artificial intelligence, when applied correctly, can significantly improve customer satisfaction and business performance, positioning companies at the forefront of innovation in the digital age.

Keywords:

Artificial Intelligence | Generative artificial intelligence | Large language models | Marketing | Intelligent chatbot

Садржај

1	УВОД.....	1
1.1	МЕТОДОЛОШКО-ХИПОТЕТИЧКИ ОКВИРИ ИСТРАЖИВАЊА.....	2
1.1.1	ПРОБЛЕМ ИСТРАЖИВАЊА.....	2
1.1.2	ПРЕДМЕТ ИСТРАЖИВАЊА.....	2
1.1.3	ЦИЉЕВИ И ЗАДАЦИ ИСТРАЖИВАЊА.....	3
1.1.4	ХИПОТЕЗЕ ИСТРАЖИВАЊА.....	5
1.1.5	МЕТОДОЛОГИЈА ИСТРАЖИВАЊА.....	7
1.2	СТРУКТУРА РАДА.....	7
2	ПРЕГЛЕД (ГЕНЕРАТИВНЕ) ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ И ВЕЛИКИХ ЈЕЗИЧКИХ МОДЕЛА.....	8
2.1	ДЕФИНИЦИЈА И ИСТОРИЈСКИ РАЗВОЈ.....	8
2.2	КЉУЧНЕ ОБЛАСТИ РАЗВОЈА ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ.....	10
2.2.1	КОМПЈУТЕРСКИ ВИД.....	11
2.2.2	РАСПЛИНУТА ЛОГИКА.....	11
2.2.3	ЕКСПЕРТСКИ СИСТЕМИ.....	11
2.2.4	РОБОТИКА.....	11
2.2.5	МАШИНСКО УЧЕЊЕ.....	12
2.2.6	НЕУРОНСКЕ МРЕЖЕ/ДУБОКО УЧЕЊЕ.....	12
2.2.7	ОБРАДА ПРИРОДНОГ ЈЕЗИКА.....	12
2.3	КРАТАК ПРЕГЛЕД МАШИНСКОГ УЧЕЊА.....	13
2.3.1	НАДГЛЕДАНО И НЕНАДГЛЕДАНО УЧЕЊЕ.....	13
2.3.2	УЧЕЊЕ СА ПОТКРЕПЉЕЊЕМ.....	14
2.4	УВОД У ГЕНЕРАТИВНУ ВЕШТАЧКУ ИНТЕЛИГЕНЦИЈУ.....	14
2.4.1	ДЕФИНИЦИЈА ГЕНЕРАТИВНИХ МОДЕЛА.....	14
2.4.2	ПРИМЕНА ГЕНЕРАТИВНИХ МОДЕЛА У РАЗЛИЧИТИМ ИНДУСТРИЈАМА.....	16
2.4.3	ТЕХНОЛОШКИ АСПЕКТИ ГЕНЕРАТИВНЕ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ.....	17
2.4.4	ОСНОВНИ ПРИНЦИПИ И МЕХАНИЗМИ РАДА.....	17
2.5	ДЕТАЉАН ПРЕГЛЕД ВЕЛИКИХ ЈЕЗИЧКИХ МОДЕЛА.....	19
2.6	СОФТВЕРСКИ АЛАТИ И ПЛАТФОРМЕ ЗА РАЗВОЈ АПЛИКАЦИЈА БАЗИРАНИХ НА ВЕШТАЧКОЈ ИНТЕЛИГЕНЦИЈИ.....	21
2.7	ЕКОНОМСКИ АСПЕКТИ РАЗВОЈА (ГЕНЕРАТИВНЕ) ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ.....	23
2.7.1	ИНТЕГРАЦИЈА (ГЕНЕРАТИВНЕ) ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ У СТРАТЕГИЈЕ ИНТЕЛИГЕНТНОГ МАРКЕТИНГА.....	23
2.7.2	АНАЛИЗА УТИЦАЈА ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ НА ПОТРОШАЧКО ПОНАШАЊЕ И ОДЛУЧИВАЊЕ.....	24
3	ПРЕГЛЕД ПОСТОЈЕЋИХ НАУЧНИХ РАДОВА.....	26

4	МЕТОДОЛОГИЈА ИСТРАЖИВАЊА	30
4.1	ПРИКУПЉАЊЕ ПОДАТАКА	30
4.2	РЕЗУЛТАТИ ИСТРАЖИВАЊА.....	31
4.3	АНАЛИЗА РЕЗУЛТАТА ЕМПИРИЈСКОГ ИСТРАЖИВАЊА	33
5	РАЗВОЈ И ИМПЛЕМЕНТАЦИЈА ИНТЕЛИГЕНТНОГ ЧЕТБОТА БАЗИРАНОГ НА ВЕЛИКИМ ЈЕЗИЧКИМ МОДЕЛИМА У ОКВИРУ АРХИТЕКТУРЕ ПРОШИРЕНЕ ГЕНЕРАЦИЈЕ.....	44
5.1	ПРЕГЛЕД СТРУКТУРЕ ПРОЈЕКТА.....	44
5.2	ТЕХНИЧКА ДОКУМЕНТАЦИЈА ФУНКЦИОНАЛНОСТИ И СТРУКТУРЕ ПРОЈЕКТА	46
5.2.1	ПРОЦЕС ЧАНКОВАЊА.....	47
5.2.2	ГЕНЕРИСАЊЕ УРЕЂЕНИХ ПАРОВА ПИТАЊЕ-ОДГОВОР.....	48
5.2.3	ОБУЧАВАЊЕ МОДЕЛА	49
5.2.4	ЕВАЛУАЦИЈА ОБУЧЕНОГ МОДЕЛА.....	51
5.2.5	ПРОЦЕС ИМПЛЕМЕНТАЦИЈЕ АРХИТЕКТУРЕ ПРОШИРЕНЕ ГЕНЕРАЦИЈЕ И ПРОЦЕСА САЖИМАЊА ОДГОВОРА КОРИШЋЕЊЕМ ВЕЛИКИХ ЈЕЗИЧКИХ МОДЕЛА	52
5.2.6	ИМПЛЕМЕНТАЦИЈА СИСТЕМА НА ВЕБ-СЕРВЕРУ	54
5.2.7	ГРАФИЧКИ КОРИСНИЧКИ ИНТЕРФЕЈС	55
5.2.8	КОНТЕЈНЕРИЗАЦИЈА СИСТЕМА	55
5.3	ТЕСТИРАЊЕ И РАД АПЛИКАЦИЈЕ.....	56
	ЗАКЉУЧАК	59
	ЛИТЕРАТУРА.....	62
	ПРИЛОГ 1: СТУДИЈСКО-ИСТРАЖИВАЧКА АНКЕТА.....	65
	ПРИЛОГ 2: ИЗВОРНИ КОД ПРОЈЕКТА	74

Списак табела

Табела 1 - Преглед постојећих истраживања.....	26
Табела 2 - Полна структура	31
Табела 3 - Старосна структура.....	32
Табела 4 - Структура према радном стажу	33

Списак илустрација

Илустрација 1 - Полна структура	31
Илустрација 2 - Старосна структура	32
Илустрација 3 - Структура према радном стажу	33
Илустрација 4 - Да ли сте упознати са појмом (генеративна) вештачка интелигенција и њеним алатима?	34
Илустрација 5 - Са којим алатима (генеративне) вештачке интелигенције сте упознати?	35
Илустрација 6 - Оцените своје разумевање (генеративне) вештачке интелигенције?	35
Илустрација 7 - Користите ли неки облик (генеративне) вештачке интелигенције у свакодневном раду?	36
Илустрација 8 - Молимо Вас да опишете на који начин користите (генеративну) вештачку интелигенцију у свакодневном раду?	36
Илустрација 9 - Како сматрате да (генеративна) вештачка интелигенција може унапредити Ваш рад?	37
Илустрација 10 - Које пословне праксе у Вашем радном окружењу би (генеративна) вештачка интелигенција могла олакшати?.....	38
Илустрација 11 - Које предности би очекивали од развоја и имплементације (генеративне) вештачке интелигенције у Вашем раду?.....	38
Илустрација 12 - Да ли сте забринути због употребе (генеративне) вештачке интелигенције у Вашем раду?.....	39
Илустрација 13 - Разлози за забринутост у коришћењу (генеративне) вештачке интелигенције.....	39
Илустрација 14 - Да ли сте упознати са концептом интелигентног четбота као алата за подршку у комуникацији и раду?	40
Илустрација 15 - Колико сте заинтересовани за коришћење интелигентног четбота у Вашем радном окружењу?	41
Илустрација 16 - Које потенцијалне предности видите у коришћењу интелигентног четбота у Вашем раду?	41

Илустрација 17 - Које изазове предвиђате при имплементацији интелигентног четбота у Ваше радно окружење?	42
Илустрација 18 - Да ли имате неке специфичне предлоге или функционалности које бисте желели да видите у интелигентном четботу?	43
Илустрација 19 - На који начин сматрате да би интелигентни четбот могао утицати на квалитет Вашег рада?	43
Илустрација 20 - Матрица конфузије.....	52
Илустрација 21 - Покретање апликације четбота	56
Илустрација 22 - Приказ одговора четбота	57
Илустрација 23 - Историја разговора.....	57
Илустрација 24 - Приказ када одговор није пронађен.....	58
Илустрација 25 - Приказ грешке	58
Илустрација 26 - Изворни код пројекта.....	74
Илустрација 27 - Изворни код пројекта.....	75
Илустрација 28 - Изворни код пројекта.....	76
Илустрација 29 - Изворни код пројекта.....	77
Илустрација 30 - Изворни код пројекта.....	78
Илустрација 31 - Изворни код пројекта.....	79
Илустрација 32 - Изворни код пројекта.....	80

1 Увод

Појава генеративне вештачке интелигенције представља револуционарни скок у начину пословања компанија, посебно у домену маркетинга и интеракције са купцима. Како се дигитална трансформација убрзава, потреба за иновативним решењима која могу да задовоље растуће захтеве савремених потрошача никада није била критичнија. Мастер теза се бави применом генеративне вештачке интелигенције и великих језичких модела са циљем да истражи трансформативни потенцијал ових технологија у повећању ангажовања корисника и оперативне ефикасности.

Генеративна вештачка интелигенција, коју карактерише њена способност да креира нови садржај и одговоре који су контекстуално релевантни и персонализовани, нуди нову парадигму у служби за кориснике. За разлику од традиционалних система вештачке интелигенције који се ослањају на унапред дефинисана правила и скупове података, генеративни модели могу да уче из огромних количина података и генеришу нове резултате, што их чини посебно погодним за динамична и непредвидива окружења као што је служба за кориснике. Фокус овог истраживања је на развоју и имплементацији интелигентног четбота, који користи архитектуру *Retrieval-Augmented Generation (RAG)*. Четбот није само алат за аутоматизацију упита купаца, већ софистицирани систем способан да разуме и одговори на сложене потребе корисника, чиме се побољшава целокупно корисничко искуство.

У средишту овог истраживања је истраживање великих језичких модела као што су *GPT* и *Llama*, који су показали изузетне способности у разумевању и генерисању текста налик човеку. Ови модели, обучени на огромним корпусима текстуалних података, могу да обављају широк спектар задатака, од одговарања на упите купаца до генерисања маркетиншког садржаја, што их чини непроцењивим алатима у дигиталном добу. Теза се такође бави *RAG* архитектуром, новим приступом који комбинује технике засноване на проналажењу са генеративним моделима како би се произвели тачнији и релевантнији одговори. Ова архитектура је посебно корисна у сценаријима где информације морају бити и прецизне и контекстуално прикладне, као што је интеракција са корисничким сервисом.

Теза истражује шире импликације интеграције вештачке интелигенције у пословању, посебно на начин на који може да подстакне иновације и конкурентску предност. Аутоматизацијом рутинских задатака и пружањем дубоких увида у понашање купаца, генеративна вештачка интелигенција омогућава предузећима да раде ефикасније и ефективније. Интелигентни четбот је замишљен као камен темељац дигиталне стратегије компаније, који помаже да се поједностави пословање, смање трошкови и побољша лојалност купаца.

1.1 Методолошко-хипотетички оквири истраживања

1.1.1 Проблем истраживања

Брзи напредак генеративне вештачке интелигенције, посебно у контексту великих језичких модела (*Large Language Models, LLMs*), донео је значајне промене у начину на који се свакодневни радни процеси спроводе у савременим пословним окружењима. Ова технолошка промена је посебно очигледна у индустријама које су у процесу дигиталне трансформације, где су генеративни алати све више интегрисани у свакодневне операције. Међутим, ова брза интеграција је такође увела низ изазова и неизвесности. Срж истраживачког проблема лежи у разумевању двоструке природе ове трансформације: с једне стране, потенцијала генеративне вештачке интелигенције да побољша продуктивност и ефикасност, а с друге стране, сложености које она уводи, као што је потреба за новим сетовима вештина, потенцијал за померање посла и утицај на добробит запослених.

Компаније усвајају решења заснована на вештачкој интелигенцији како би поједноставили пословање, побољшали корисничку услугу и побољшали процесе доношења одлука. Иако ова унапређења нуде бројне предности, постоји и све већа забринутост око тога како ове технологије утичу на свакодневне радне рутине запослених. Да ли ови алати поједностављују задатке или стварају додатне слојеве сложености? Да ли су запослени у стању да ефикасно искористе ове алате или се боре са прилагођавањем на нове токове рада вођене вештачком интелигенцијом?

Проблем истраживања, се може артикулисати на следећи начин: „У којој мери интеграција генеративне вештачке интелигенције у свакодневне радне процесе побољшава или отежава продуктивност запослених и доношење одлука у свакодневном пословању?“ Ова изјава о проблему покреће истраживање позитивних утицаја и потенцијалних недостатака усвајања вештачке интелигенције у свакодневним радним активностима, са циљем да пружи уравнотежен поглед на импликације генеративне вештачке интелигенције на модерне пословне праксе.

Разумевање овог проблема је кључно из неколико разлога. Прво, бави се празнином у постојећим истраживањима која се често више фокусирају на технолошке могућности вештачке интелигенције, а не на њене импликације у стварном свету за запослене. Друго, налази овог истраживања могли би да информишу о бољим стратегијама за имплементацију вештачке интелигенције, обезбеђујући да се такве технологије искористе у пуном потенцијалу, док се негативни утицаји на запослене минимизирају. Како генеративна вештачка интелигенција наставља да се развија, од суштинске је важности да се континуирано процењује њен утицај на радне процесе како би се осигурало да позитивно доприноси пословним циљевима и задовољству запослених.

1.1.2 Предмет истраживања

Предмет овог истраживања фокусиран је на специфичне начине примене генеративне вештачке интелигенције у свакодневним радним процесима у оквиру савременог пословања.

Ова студија се бави применом алата генеративне вештачке интелигенције у стварном свету као што су модели великих језика и њихов утицај на свакодневне задатке запослених. Испитујући како су ова решења вођена вештачком интелигенцијом интегрисана у рутинске радне активности, истраживање има за циљ да открије шире импликације усвајања вештачке интелигенције у професионалном окружењу.

Последњих година, виђен је значајан помак ка дигиталној трансформацији, вођен потребом да компаније остану конкурентни у технолошком свету који се брзо мења. Као део ове трансформације, уведени су генеративни алати за побољшање различитих аспеката пословних операција, од корисничке подршке, до интерних процеса доношења одлука. Ови алати су дизајнирани да аутоматизују задатке који се понављају, помажу у решавању сложених проблема и пружају увид на основу великих скупова података. Међутим, степен до којег су ови алати ефикасно интегрисани у свакодневни радни живот запослених и даље је недовољно истражен. Кључни аспекти предмета истраживања су:

- **Интеграција вештачке интелигенције:** како су генеративни алати вештачке интелигенције интегрисани у постојеће токове рада, укључујући типове задатака које је вештачка интелигенција аутоматизовала или побољшала.
- **Интеракција запослених:** начини на које запослени ступају у интеракцију са алаткама вештачке интелигенције, укључујући њихово разумевање ових алата, једноставност коришћења и било какву обуку која се пружа да би се олакшала ова интеракција.
- **Утицај на радне процесе:** директни ефекти вештачке интелигенције на продуктивност, доношење одлука и задовољство послом, као и све нежељене последице, као што су повећано оптерећење или страх од промене посла.
- **Изазови усвајања:** идентификовање препрека успешном усвајању вештачке интелигенције, као што су отпор променама, недостатак техничких вештина или неадекватни системи подршке.

Предмет овог истраживања је од суштинског значаја за разумевање стварног утицаја вештачке интелигенције на пословне операције и искуства запослених. Фокусирајући се на савремено пословање, студија пружа детаљно испитивање улоге вештачке интелигенције у специфичном контексту пословања, нудећи увиде који би могли да буду применљиви на све секторе који пролазе кроз дигиталне трансформације. Налази овог истраживања могу помоћи да се информишу о будућим имплементацијама вештачке интелигенције, обезбеђујући да се ове технологије примењују на начине који максимизирају користи уз ублажавање потенцијалних недостатака.

1.1.3 Циљеви и задаци истраживања

Примарни циљ овог истраживања је процена утицаја и примене генеративне вештачке интелигенције на свакодневне радне процесе запослених у савременим пословним окружењима. Конкретно, ова студија има за циљ да разуме како ове технологије вештачке

интелигенције, посебно модели великих језика, преобликују радне праксе, повећавају продуктивност и утичу на процесе доношења одлука. Фокус је на идентификовању позитивних утицаја и потенцијалних изазова који долазе са интеграцијом генеративне вештачке интелигенције у свакодневне радне рутине. Специфични циљеви истраживања су:

1. **Процена коришћења вештачке интелигенције:** одредити у којој мери се генеративни алати користе у свакодневним операцијама од стране запослених у савременим пословним окружењима. Ово укључује идентификовање типова задатака које најчешће подржава или аутоматизује и учесталост коришћења алата за вештачку интелигенцију.
2. **Процена утицаја на продуктивност:** анализирати како усвајање генеративних алата утиче на продуктивност запослених. Овај циљ укључује истраживање да ли интеграција вештачке интелигенције доводи до уштеде времена, побољшане тачности задатака или побољшане ефикасности у обављању свакодневних радних активности.
3. **Истраживање побољшања доношења одлука:** испитивање улоге вештачке интелигенције у подршци или побољшању процеса доношења одлука у савременим пословним контекстима. Ово укључује разумевање начина на који се алати користе за пружање увида, предвиђање исхода или давање препорука које помажу запосленима у доношењу одлука на основу информација.
4. **Идентификовање изазова:** идентификовати изазове са којима се запослени суочавају када интегришу генеративну вештачку интелигенцију у своје радне процесе. Овај циљ се бави потенцијалним питањима као што су отпор променама, недостатак техничке експертизе и сложеност интеракције са системима вештачке интелигенције.
5. **Разумевање перцепције запослених:** прикупити увид у то како запослени перципирају утицај вештачке интелигенције на њихов рад, укључујући њихове ставове о предностима, недостацима и општој ефикасности генеративних алата у побољшању њиховог радног учинка.

Да би се постигли ови циљеви, истраживање је структурирано око следећих кључних задатака:

1. **Преглед литературе:** спровести свеобухватан преглед постојећих истраживања о примени генеративне вештачке интелигенције у пословним окружењима, фокусирајући се на њен утицај на продуктивност, доношење одлука и радне процесе. Овај задатак ће пружити теоријску основу за разумевање тренутног стања усвајања генеративне вештачке интелигенције на радном месту.
2. **Прикупљање података:** анкете ће се користити за прикупљање података. Ови подаци ће бити критични у процени начина на који се генеративни алати користе и њиховог утицаја на свакодневне радне процесе.

3. **Анализа података:** анализирање прикупљених података користећи одговарајуће статистичке методе да би се идентификовали трендове, корелације и обрасци у коришћењу тих алата. Ова анализа ће помоћи да се квантификује утицај вештачке интелигенције на продуктивност и доношење одлука и истакне изазове са којима се суочавају запослени.
4. **Синтеза налаза:** интегришите налазе из анализе података са увидима из прегледа литературе да бисте извели свеобухватне закључке о улози генеративне вештачке интелигенције на савременим радним местима. Ова синтеза ће се бавити и предностима и ограничењима усвајања вештачке интелигенције.
5. **Препоруке:** на основу налаза истраживања, развити практичне препоруке за предузећа која желе да интегришу генеративну вештачку интелигенцију у своје радне процесе. Ове препоруке ће се фокусирати на максимизирање предности вештачке интелигенције док се баве идентификованим изазовима како би се обезбедио несметан и ефикасан процес усвајања.

Циљеви и задаци истраживања су дизајнирани да пруже темељно разумевање како генеративна вештачка интелигенција утиче на свакодневне радне процесе у савременим пословним окружењима. Фокусирајући се на ове специфичне области, истраживање има за циљ да допринесе драгоценим увидима који могу да усмере будућу имплементацију вештачке интелигенције, обезбеђујући да се ове технологије искористе у свом пуном потенцијалу, а минимизирају све негативне утицаје на запослене.

1.1.4 Хипотезе истраживања

Истраживачке хипотезе су формулисане да тестирају специфичне аспекте утицаја и примене генеративне вештачке интелигенције у свакодневним радним процесима у савременим пословним окружењима. Ове хипотезе служе као основа за емпиријско истраживање, усмеравајући процесе прикупљања података и анализе како би се утврдило да ли су очекивани исходи у складу са стварним искуствима запослених у интеракцији са генеративним алатима. На основу циљева истраживања и прегледа литературе, развијене су следеће хипотезе:

1. **Хипотеза 1 (X1):** интеграција генеративне вештачке интелигенције у свакодневне радне процесе значајно повећава продуктивност запослених.
Образложење: Генеративни алати су дизајнирани да аутоматизују рутинске задатке, смање време потребно за сложене процесе и пружају увиде који побољшавају доношење одлука. Стога се претпоставља да ће њихова интеграција довести до приметних побољшања укупне продуктивности.
2. **Хипотеза 2 (X2):** запослени са вишим нивоима писмености имају веће користи од интеграције вештачке интелигенције у поређењу са онима са нижом писменошћу у вези са вештачком интелигенцијом.

Образложење: ефикасност генеративних алата често зависи од способности корисника да их разуме и ефикасно комуницира са њима. Запослени који су боље упознати са концептима и алатима вештачке интелигенције вероватно ће ефикасније користити ове технологије, што ће довести до бољих резултата.

3. **Хипотеза 3 (Х3):** Употреба генеративних алата у процесима доношења одлука доводи до тачнијих и правовременијих одлука у савременим пословним окружењима.

Образложење: алати генеративне вештачке интелигенције, посебно они засновани на великим језичким моделима, способни су да анализирају велике скупове података и дају препоруке засноване на подацима. Очекује се да ће ова способност побољшати тачност и брзину доношења одлука, чиме ће се побољшати пословни резултати.

4. **Хипотеза 4 (Х4):** Интеграција генеративне вештачке интелигенције побољшава креативност запослених у извршавању задатака, што доводи до иновативнијих решења.

Образложење: генеративни алати могу помоћи у размишљању и генерисању нових идеја, чиме се подстиче креативност. Запослени који користе алате генеративне вештачке интелигенције могу се наћи у стању да приступе проблемима из нових углова и развију иновативнија решења.

5. **Хипотеза 5 (Х5):** Примена генеративних алата наилази на отпор запослених, посебно због забринутости око сигурности посла и сложености система вештачке интелигенције.

Образложење: упркос потенцијалним предностима вештачке интелигенције, често постоји забринутост у вези са измештањем посла и кривом учења која је повезана са новим технологијама. Ове забринутости могу довести до отпора међу запосленима, што може омести успешно усвајање тих алата.

Ове хипотезе су дизајниране да тестирају кључне аспекте утицаја вештачке интелигенције на свакодневне радне процесе, укључујући продуктивност, доношење одлука, креативност и изазове повезане са усвајањем вештачке интелигенције. Потврђивањем или побијањем ових хипотеза кроз емпиријску анализу, истраживање има за циљ да пружи свеобухватно разумевање практичних импликација интеграције на радном месту.

Да би се тестирале ове хипотезе, подаци ће бити анализирани коришћењем статистичких метода како би се утврдила снага и правац односа између употребе генеративне вештачке интелигенције и исхода од интереса (продуктивност и доношење одлука). Анализа ће се фокусирати на идентификацију значајних корелација и образаца који подржавају или оспоравају предложене хипотезе.

Истраживачке хипотезе пружају јасан оквир за истраживање утицаја генеративне вештачке интелигенције на процесе рада. Систематским тестирањем ових хипотеза, истраживање ће допринети драгоценим увидима у то како технологије вештачке интелигенције трансформишу савремено радно место, како позитивно тако и негативно.

1.1.5 Методологија истраживања

Методологија истраживања ослања се на податке прикупљене путем онлајн анкете, дистрибуиране преко *Google Forms* алата. Анкета је подељена путем *LinkedIn* мреже и на другим веб сајтовима за анкете, са циљем да се испитају искуства професионалаца у различитим индустријама у вези са применом генеративне вештачке интелигенције у њиховим свакодневним радним процесима. Истраживала је аспекте као што су учесталост коришћења ових алата, њихов утицај на продуктивност, доношење одлука, креативност и изазове у примени.

Анкета је осмишљена да обухвати квантитативне и квалитативне податке, користећи затворена питања за специфичне податке и отворена питања за дубље увиде. Добијено је 118 одговора, чиме је омогућена анализа употребе (генеративне) вештачке интелигенције у различитим радним контекстима. Подаци су анализирани квантитативним методама, укључујући дескриптивну статистику, као и квалитативним анализама, попут тематске анализе за откривање дубљих образаца у одговорима испитаника.

1.2 Структура рада

Рад је структуриран у неколико кључних поглавља, почевши од проблема и обима истраживања, које се усредсређује на разумевање како генеративна вештачка интелигенција може да побољша пословне перформансе док уводи изазове као што су потреба за новим сетовима вештина и њихов утицај на запослене. Након тога, теоријска основа покрива историјски развој вештачке интелигенције и унапређење генеративних модела, наглашавајући њихову улогу у трансформацији индустрија, укључујући и маркетинг индустрију.

Затим, теза представља опширан преглед постојећих научних радова и технолошких аспеката, посебно фокусирајући се на моделе као што су *GPT* и *Transformers* архитектуре. Поглавље анализира и техничке и стратешке импликације, нудећи увид у то како се ови алати могу ефикасно скалирати и интегрисати у маркетиншке операције.

Истраживање је засновано на онлајн анкети, прикупљајући податке од експерата из различитих индустрија. Овај део укључује детаљну анализу како генеративни алати утичу на свакодневне радне процесе, продуктивност, доношење одлука и задовољство запослених. Методологија истраживања је пажљиво осмишљена за прикупљање квалитативних и квантитативних података, пружајући свеобухватан увид у предности и изазове са којима се суочавају током интеграције генеративне вештачке интелигенције.

Завршни део рада представља пројекат развоја интелигентног четбота. Пројекат укључује дизајнирање, имплементацију и тестирање апликације која има за циљ да побољша перформансе улога које су окренуте клијентима у телекомуникационој компанији.

У закључку, мастер теза показује да генеративна вештачка интелигенција, када се примењује промишљено, има потенцијал да значајно унапреди маркетиншке процесе, задовољство купаца и оперативну ефикасност, док се истовремено бави техничким, економским и стратешким изазовима његовог усвајања.

2 Преглед (генеративне) вештачке интелигенције и великих језичких модела

2.1 Дефиниција и историјски развој

IBM даје следећу дефиницију вештачке интелигенције¹:

„Вештачка интелигенција (*Artificial Intelligence, AI*) је технологија која омогућава рачунарима и машинама да симулирају људско учење, разумевање, решавање проблема, доношење одлука, креативност и аутономију.“ (Stryker & Kavlakoglu, 2024)

Ова дефиниција има свој историјски корен на самим почецима развоја вештачке интелигенције. Британски математичар *Alan Turing* је 1936. године доказао је да је могућ универзални калкулатор – данас познат као Турингова машина. Турингов главни увид био је да је таква машина способна решити сваки проблем све док се он може представити и решити алгоритмом. Преведено на подручје људске интелигенције, то значи да ако се когнитивни процес може алгоритамски раставити на добро дефинисане појединачне кораке, тада се тај процес може извршити на једној машини. (Repi, V. V. R., et al., 2024)

Људи су од давнина маштали о томе да машине почну да уче и решавају људске интелектуалне задатке. Атрактивност таквих машина била је велика, па се посезало и за преварама попут „Турчина“, машине која је наводно могла да игра шај и побеђује људе. Данас се зна да је у „Турчину“ био сакривен човек, шахиста. Али, временом, корак по корак, постављане су основе за то да се такве машине заиста направе. *Kaynak* (Kaynak, 2021) и *Repi* (Repi, V. V. R., et al., 2024) дају преглед најважнијих етапа на том путу:

1666. Математичар и филозоф *Gottfried Leibniz* објављује *Dissertatio de arte combinatoria* (О комбинаторној уметности) и предлаже абецеду људске мисли, тврдећи да све идеје нису ништа друго него комбинације малог броја једноставних концепата.

1763. *Thomas Bayes* развија оквир за размишљање о вероватноћи догађаја. Бајесово закључивање постаће кључни концепт у машинском учењу.

1854. *George Boole* износи став да се логичко размишљање може спроводити систематски на исти начин као и решавање система једначина.

1914. Шпански инжењер *Leonardo Torres Quevedo* демонстрира прву машину за играње шаха, способну за завршницу краља и топа против краља без људске интервенције.

1943. *Warren Sturgis McCulloch* и *Walter Pitts* објавили су рад *A Logical Calculus of the Ideas Immanent in Nervous Activity* (Логички рачун идеја иманентних нервној активности) у часопису *Bulletin of Mathematical Biophysics*. Овај утицајни рад, у којем су расправљали о

¹ Овде се наводе веома скраћене тачке у развоју које се сматрају најважнијим, нарочито оне на почецима утемељења концепта вештачке интелигенције. Актуелна решења биће разматрана у наставку.

идеализованим, поједностављеним мрежама вештачких неурона који би могли изводити једноставне логичке функције, послужиће као инспирација за неуронске мреже, на којима ће се касније заснивати дубоко учење рачунара.

1950. *Alan Turing* објављује рад *Computing Machinery and Intelligence* (Рачунарска машина и интелигенција) у којем предлаже „игру опонашања“ која ће касније постати позната као „Турингов тест“.

1951. *Marvin Minsky* и *Din Edmonds* граде *SNARC (Stochastic Neural Analog Reinforcement Calculator)*, прву вештачку неуралну мрежу, користећи 3000 вакуумских цеви за симулацију мреже од 40 неурона.

1952. *Arthur Samuel* развија први рачунарски програм за играње шаха и први рачунарски програм који самостално учи.

31. августа 1955. по први пут се појављује израз „вештачка интелигенција“ скован у предлогу за двомесечну студију вештачке интелигенције коју су поднели *John McCarthy*, *Marvin Minsky*, *Nathaniel Rochester* и *Claude Elwood Shannon*. Радионице које су одржане годину дана касније, 1956, сматрају се службеним датумом рађања ове нове области.

Децембра 1955. *Herbert Simon* и *Allen Newell* развили су *Logic Theorist*, први програм вештачке интелигенције који је доказао 38 од прве 52 теореме у *Whitehead* и *Russell* књизи *Principia Mathematica*.

1957. *Frank Rosenblatt* развија *Perceptron*, рану вештачку неуралну мрежу која омогућава препознавање узорака на темељу двослојних рачунарских мрежа за учење.

1958. *John McCarthy* развио је програмски језик *Lisp* који је постао најпопуларнији програмски језик коришћен у истраживању вештачке интелигенције.

1959. *Arthur Samuel* сковао је израз "машинско учење", пишући о програмирању рачунара како би он научио играти бољу партију шаха него што би особа која је написала програм могла играти.

1959. *Oliver Selfridge* објављује рад *Pandemonium*: парадигма за учење у *Proceedings of the Symposium on the Mechanization of Thought Processes*, у којем описује модел за процес помоћу којег рачунари могу препознати обрасце који нису претходно одређени.

1959. *John McCarthy* објављује чланак Програми са здравим разумом у *Proceedings of the Symposium on the Mechanization of Thought Processes*, у којем описује *Suggestion Taker*, програм за решавање проблема манипулисањем реченицама у формалном језику с крајњим циљем стварања програма који ће учити из искуства једнако ефикасно као и људи.

1964. *Daniel G. Bobrow* завршио је своју докторску дисертацију на *MIT* под насловом Унос природног језика за рачунарски систем решавања проблема и развио *STUDENT*, рачунарски програм за разумевање природног језика.

1965. *Joseph Weizenbaum* развија интерактивни програм *ELIZA2*, који води дијалоге на енглеском језику о било којој теми. Аутор је био изненађен бројем људи који су рачунарским програмима приписали људске осећаје.

1968. *Terry Winograd* развија *SHRDLU*, рани рачунарски програм за разумевање природног језика.

1969. *Arthur E. Bryson* описује ширење унатраг као вишестепену методу оптимизације динамичког система. Овај алгоритам учења за вишеслојне вештачке неуронске мреже значајно је допринео успеху дубоког учења после 2000. године, након што је снага рачунара довољно напредовала да се прилагоди потребама обуке великих мрежа.

1995. *Richard Wallace* развија *ALICE (Artificial Linguistic Internet Computer Entity)* четбот, инспирисан програмом *ELIZA*, али с додатком прикупљања узорака података природног језика у великом обиму, што је омогућено појавом веба.

1997. *Deep Blue* постаје први рачунарски програм за играње шаха који је победио актуалног светског првака у шаху.

2009. Компјутерски научници у *Intelligent Information Laboratory* на универзитету *Northwestern* развијају *Stats Monkey*, програм који пише спортске вести без људске интервенције.

2011. Конволуциона неуронска мрежа побеђује на немачком такмичењу у препознавању саобраћајних знакова са 99,46% тачности (у поређењу са људима који постижу тачност од 99,22%).

2.2 Кључне области развоја вештачке интелигенције

Већина аутора издваја 4 - 7 кључних области развоја вештачке интелигенције. У раду преузета је класификација коју даје *Marisa Garanhel (Garanhel, 2023)*. По мишљењу овог аутора седам кључних области развоја јесу:

1. Компјутерски вид,
2. Расплинута логика (*Fuzzy logic*),
3. Експертски системи,
4. Роботика,
5. Машинско учење,
6. Неуронске мреже/дубоко учење и
7. Обрада природног језика.

² Треба нагласити да је *ELIZA* била више згодна досетка него ма каква примена вештачке интелигенције. Значај овог програма највише је у томе што је скренуо пажњу на то како вештачка интелигенција може бити предмет приписивања антропоморфних карактеристика које нема од стране људи.

2.2.1 Компјутерски вид

Компјутерски вид има за циљ да развије технике које помажу рачунарима да виде и разумеју дигиталне слике и видео записе. Примена модела машинског учења на слике омогућава рачунарима да идентификују објекте, лица, људе, животиње итд. Алгоритамски модели помажу рачунарима да уче о контексту визуелних података, а са довољно података који се уносе кроз модел, рачунари могу сами себе да науче да разликују једну слику од друге. Конволуциона неуронска мрежа ради заједно са моделом који разбија слике на пикселе, дајући им ознаке. Неуронска мрежа затим користи те ознаке за обављање конволуције, што је математичка операција са две функције које производе трећу функцију која прави предвиђања о томе шта рачунар „види“. Рачунарски вид има примену у различитим областима, као што су праћење објеката, класификација слика и препознавање лица.

2.2.2 Расплинута логика

Расплинута логика је погодан начин мапирања улазних и излазних података. Између те две крајње тачке налази се процесор података. Могу се мапирати најразличитији подаци. За разлику од Бајесове формуле, *Fuzzy* логика омогућава много шири скуп улазних података и подесна је онда када не знамо тачно неке важне параметре, јер они тек треба да се реализују у будућности. Показало се да је у таквим случајевима, *Fuzzy* логика врло добар приступ који може дати значајне предикције. (Fuzzy Logic Toolbox, 2002)

Овај метод узима у обзир све постојеће могућности између дигиталних вредности „да“ и „не“. Поједностављено, мери степен до ког је хипотеза тачна. Ова грану вештачке интелигенције користи се за разматрање неизвесних тема. То је згодан и флексибилан начин имплементације техника машинског учења и логичког копирања начина на који људи размишљају. Овај приступ се користи нпр. за аутоматско спречавање несрећа током вожње аутомобила. (Garanhel, 2023)

2.2.3 Експертски системи

Експертски систем је програм специјализован за појединачни задатак, баш као људски експерт. Ови системи су дизајнирани да решавају замршене проблеме са способностима доношења одлука сличних људима. Они користе скуп правила, названих правила закључивања, која за њих дефинише база знања која се снабдева подацима. Користећи ако-онда логичке низове, они могу решити сложена питања и помоћи у управљању информацијама, постављању медицинских дијагноза, анализи кредита и многим другим областима. Пример експертског система је *CaDeT*, систем дијагностичке подршке који може помоћи медицинским радницима откривањем рака у раним фазама. (Garanhel, 2023)

2.2.4 Роботика

Роботи су програмиране машине које могу аутоматски да извршавају сложене серије акција. Роботи помажу људима у заморним и понављајућим задацима. Посебно роботи са

вештачком интелигенцијом могу самостално деловати нпр. на задацима у свемиру. „Sofia, робот који је развио *Hanson Robotics*, ради кроз комбинацију вештачке интелигенције и неуронских мрежа. Она препознаје људска лица и разуме емоције и гестове – и може чак да комуницира са људима.“ (Garanhel, 2023)

2.2.5 Машинско учење

Машинско учење је способност машина да аутоматски уче из података и алгоритама и једна је од најзахтевнијих грана вештачке интелигенције. „Машинско учење побољшава перформансе користећи прошла искуства и може доносити одлуке без посебног програмирања за то. Процес почиње прикупљањем историјских података, попут инструкција и директног искуства, тако да се могу изградити логички модели за будуће закључке. Тачност излаза зависи од величине података – већа количина података ће изградити бољи модел, што заузврат повећава његову тачност.“

2.2.6 Неуронске мреже/дубоко учење

Неуронске мреже су познате и као вештачке неуронске мреже (*Artificial Neural Networks, ANN*) или симулиране неуронске мреже (*Simulated Neural Networks, SNN*). Налазе се у језгру алгоритама и инспирисане су људским мозгом: копирају како биолошки неурони сигнализирају једни другима. *ANN* имају слојеве чворова – који се састоје од улазног слоја, једног или више скривених слојева и излазног слоја. Сваки чвор, који се такође назива вештачки неурон, повезује се са другим неуронима и има придружени праг и тежину. Када је излаз појединачног чвора изнад одређене вредности прага, чвор се активира да пошаље податке следећем мрежном слоју. Неуронским мрежама су потребни подаци за обуку како би учили и побољшали тачност.

2.2.7 Обрада природног језика

Обрада природног језика омогућава рачунарима да разумеју и текст и изговорене речи попут људи. Комбинујући машинско учење, лингвистику и моделе дубоког учења, рачунари могу да обрађују људски језик у гласовним или текстуалним подацима да би разумели пуно значење, намере и осећања. У препознавању говора гласовни подаци се поуздано могу претварати у текстуалне податке. Неки случајеви употребе обраде природног језика су:

- **Виртуелни четботови:** они могу препознати контекстуалне информације како би купцима понудили боље одговоре током времена.
- **Откривање нежељене поште:** класификација текста за обраду природног језика може скенирати језик у имејловима да би открила нежељену пошту.
- **Анализа сентимената:** анализа језика који се користи на платформама друштвених медија помаже да се извуку емоције и ставови о производима.

2.3 Кратак преглед машинског учења

Последњих година сведоци смо изузетног напретка вештачке интелигенције. Област вештачке интелигенције која се назива машинско учење омогућила је рачунарима да уче из података без експлицитног програмирања, што је довело до напретка у препознавању слика, обради природног језика и системима за давање препорука. Дубоко учење, као један од облика машинског учења, још више је померило границе вештачке интелигенције, омогућавајући машинама да препознају обрасце и доносе одлуке на нивоу за који се раније мислило да је немогућ. (Stokel-Walker, 2024).

2.3.1 Надгледано и ненадгледано учење

Један од најчешће коришћених типова алгоритама машинског учења је надгледано учење. У надгледаном учењу алгоритам се обучава на означеном скупу података, где су улазни подаци упарени са исправним излазом. Алгоритам учи да мапира улазе у излазе проналажењем образаца у подацима. Овај тип алгоритма се често користи у задацима као што су класификација и регресија, где је циљ пронаћи одговарајућу ознаку или вредност на основу улазних података. (Hernandez, 2024)

Подаци за обуку састоје се од познатих (означених) улазних и излазних парова. На пример, у класификацији слика, улаз је слика, а излаз је одговарајућа ознака категорије. Модели се обучавају да дају предвиђања или класификују нове податке на темељу узорака пронађених у подацима за обуку. Модел се прилагођава подацима за обуку помоћу алгоритма надгледаног учења. Погрешка предвиђања се пореди са стварном вредношћу, а параметри модела се ажурирају како би се погрешка свела на минимум. Модел се процењује коришћењем података који нису коришћени током обуке да би се измерио опсег до којег се модел може генерализовати на нове податке. Различите врсте модела могу се користити за надгледано учење, укључујући линеарну регресију, логистичку регресију, потпорне векторске машине и вештачке неуронске мреже. (Repi, V. V. R., et al., 2024)

Квалитет и репрезентативност података за обуку врло су важни у надгледаном учењу, јер модел може учити само из информација садржаних у подацима.

Друга врста алгоритма машинског учења је ненадгледано учење. У ненадгледаном учењу алгоритам добија скуп података без икаквих ознака, а његов циљ је да пронађе обрасце и структуру у подацима. Овај тип алгоритма се често користи у задацима као што су груписање и смањење броја димензија, где је циљ груписање сличних тачака података заједно или смањење сложености података. (Hernandez, 2024)

Подаци за обуку овде немају означене улазно-излазне парове. Модел, дакле, мора пронаћи структуре или узорке у подацима без помоћи ознака. Он групише или класифицира податке у категорије или кластере на темељу постојећих сличности или образаца. Евалуација ненадгледаног учења субјективнија је од надгледаног учења јер нема очекиваних ознака с којима би се могло поредити. Модели могу произвести излаз који одражава пристрасност или субјективно тумачење података, будући да нема објективног мерења путем ознака. Почетне одлуке или иницијализација модела могу утицати на исходе ненадгледаног учења, а резултати могу варирати зависно од различитих иницијализација. Постоје различите методе у

ненадгледаном учењу, укључујући кластеризацију k -средњих вредности, факторску анализу и аутокодере у дубоком учењу. (Repí, V. V. R., et al., 2024)

Ненадгледано учење игра важну улогу у истраживању и разумевању података који немају јасне ознаке или структуру. То се може користити за откривање скривених образаца и може пружити вредне увиде у анализу података.

2.3.2 Учење са поткрепљењем

Учење са поткрепљењем (*Reinforcement learning*) је врста машинског учења где модел учи кроз интеракцију са околином. Овај модел прима повратну информацију у облику награда или казни као одговор на предузете акције, а циљ му је максимизовати акумулацију награда током времена. (Hernandez, 2024)

Модел ступа у интеракцију с околином и бира радње које сматра најприкладнијим на основу тренутних околности. Околина моделу даје повратну информацију у облику награда или казни, указујући на квалитет акције. Главни циљ модела јесте максимизовање укупног броја примљених награда током времена. Модел учи путем покушаја и погрешака, где покушава различите акције и посматра резултате. Акције предузете у једном тренутку могу утицати на будуће награде. Ово уводи концепт одгођених награда. Модел покушава идентификовати обрасце деловања који производе добре резултате и тако развија ефикасне стратегије. Изазов у *Reinforcement learning* приступу је проналажење равнотеже између истраживања (испробавање нових акција) и искоришћавања (предузимање акција за које се већ зна да доносе награде). (Repí, V. V. R., et al., 2024)

Reinforcement learning може се користити у различитим контекстима у којима модел треба да комуницира са околином и научи оптималне акције током времена. Овај метод је дао значајне резултате у разним применама, укључујући игре, роботiku и управљање ресурсима.

2.4 Увод у генеративну вештачку интелигенцију

2.4.1 Дефиниција генеративних модела

Генеративни модели су класа статистичких модела који могу да генеришу нове инстанце података, често се користе у ненадгледаном машинском учењу као средство за обављање задатака. На пример, генеративни модели се могу користити за вршење процене вероватноће, моделирање тачака података, описивање феномена у подацима или разликовање између класа на основу ових вероватноћа. У контексту мултимедије, генеративни модели се могу користити за генерисање података као што су слике и аудио. (Chen, et all, 2024)

„Генеративни модел је тип модела машинског учења који има за циљ да научи основне обрасце или дистрибуције података у циљу генерисања нових, сличних података. У суштини, то је као да подучавате рачунар да измишља сопствене податке на основу онога што је раније видео. Значај овог модела лежи у његовој способности стварања, што има огромне импликације у различитим областима, од уметности до науке.“ (Awan, 2023)

Са ширењем метода дубоког учења (*Deep Learning*) појавила се нова породица метода, названа дубоки генеративни модели, формирана је кроз комбинацију генеративних модела и дубоких неуронских мрежа. (Li & Gao, 2020)

Генеративни модели долазе у различитим облицима, сваки са јединственим приступом разумевању и генерисању података. Ево свеобухватније листе неких од најистакнутијих типова (Awan, 2023):

- **Бајесове мреже:** графички модели који представљају вероватноће везе између скупа варијабли. Они су посебно корисни у сценаријима у којима је разумевање узрочно-последичних веза кључно. На пример, у медицинској дијагнози Бајесова мрежа може помоћи у одређивању вероватноће болести с обзиром на низ симптома.
- **Дифузиони модели:** модели описују како се ствари шире или развијају током времена. Често се користе у сценаријима као што је разумевање како се гласина шири у мрежи или предвиђање ширења вируса у популацији.
- **Генеративне адверсаријске мреже:** састоје се од две неуронске мреже, генератора и дискриминатора, које су заједно обучене. Генератор покушава да произведе податке, док дискриминатор покушава да направи разлику између стварних и генерисаних података. Временом, генератор постаје толико добар да дискриминатор не може да уочи разлику. GAN су популарни у задацима генерисања слика, као што је стварање реалистичних људских лица или уметничких дела.
- **Вариацијски аутоенкодер:** тип аутоенкодера који производи компримовани приказ улазних података, а затим их декодира да би генерисао нове податке. Често се користе у задацима попут уклањања шума слике или генерисања нових слика које деле карактеристике са улазним подацима.
- **Ограничене Болцманове машине:** неуронске мреже са два слоја које могу научити дистрибуцију вероватноће преко свог скупа улаза. Коришћени су у системима препорука, као што је предлагање филмова на платформама за стриминг на основу корисничких преференција.
- **Рекурентне неуронске мреже пиксела:** ови модели генеришу слике пиксел по пиксел, користећи контекст претходних пиксела за предвиђање следећег. Посебно су корисни у задацима где је секвенцијално генерисање података кључно, као што је цртање слике линију по линију.
- **Марков ланци:** модели који предвиђају будућа стања искључиво на основу садашњег стања, не узимајући у обзир стања која су му претходила. Често се користе у генерисању текста, где се следећа реч у реченици предвиђа на основу тренутне речи.

- **Нормализација токова:** Ово су серије инверзибилних трансформација које се примењују на једноставне дистрибуције вероватноће да би се добиле сложеније дистрибуције. Корисни су у задацима где је разумевање трансформације података кључно, нпр. у финансијском моделирању.

2.4.2 Примена генеративних модела у различитим индустријама

Генеративна вештачка интелигенција је тренд који све више налази своје место у најразличитијим индустријама. *Cursera* (2024) даје преглед најкарактеристичнијих примена у данашњем времену:

- **Здравствена заштита и фармација:** генеративна вештачка интелигенција има примену за све делове здравствене и фармацеутске индустрије, од откривања и развоја нових лекова који спасавају животе до персонализовања планова лечења за појединачне пацијенте до стварања предиктивних слика за одређивање прогресије болести.
- **Оглашавање и маркетинг:** генеративна вештачка интелигенција нуди многа решења професионалцима који раде у оглашавању и маркетингу, као што је генерисање текста и слика потребних за маркетинг или проналажење нових начина за интеракцију са купцима.
- **Производња:** у производњи, професионалци могу да користе генеративну вештачку интелигенцију да траже начине за побољшање ефикасности, предвиде потребе за одржавањем пре него што се појаве проблеми, помогну инжењерима да брже креирају боље дизајне и креирају отпорнији ланац снабдевања.
- **Развој софтвера:** за тим за развој софтвера, генеративна вештачка интелигенција може да обезбеди алате за брже креирање и оптимизацију кода.
- **Финансијске услуге:** неке од примена генеративне вештачке интелигенције у индустрији финансијских услуга укључују стратегије улагања вештачке интелигенције, израду документације и праћење регулаторних промена и коришћење генеративне вештачке интелигенције као тумача за олакшавање комуникације између клијената и инвеститора.
- **Медији и забава:** медији и индустрија забаве би могли да прихвате генеративну вештачку интелигенцију на неколико начина, пре свега за генерисање јединственог садржаја. Генеративна вештачка интелигенција може помоћи у креирању и уређивању визуелног садржаја, креирању кратких видео записа о спортским догађајима и олакшати рад са системима за управљање садржајем.

2.4.3 Технолошки аспекти генеративне вештачке интелигенције

2.4.3.1 Преглед архитектуре

Transformers архитектура појавила се 2017. године кроз револуционарни истраживачки рад под насловом *Attention is All You Need* (Пажња је све што вам треба) који су написали Vaswani и сарадници. Од тада, *Transformers* архитектура је постала окосница најсавременијих NLP³ модела. (Hogan, 2023)

Пре *Transformers* архитектуре, традиционални *NLP* модели су се за руковање низовима података као што су реченице или параграфи првенствено ослањали на рекурентне неуронске мреже и конволуционе неуронске мреже. Мада су ови модели били у некој мери ефикасни, имали су и ограничења, посебно на пољу уочавања зависности унутар текста.

2.4.4 Основни принципи и механизми рада

Transformers архитектура се издваја по томе што поседује неколико кључних компоненти, разумевање којих је кључно да би се схватиле способности ове архитектуре и како она ради. Hogan (2023) издваја пет кључних компоненти *Transformers* архитектуре:

- **Механизам самопажње (*Self-Attention Mechanism*):** ово је кључни механизам *Transformers* архитектуре, оно што јој даје изузетну способност да обрађује низове података, као што су речи у реченици, тако што додељује различити степен пажње различитим елементима на основу њиховог контекста. Овај механизам процењује сваку реч коју обрађује у контексту осталих речи, целе реченице и околних реченица, па је нпр. способан да пронађе на шта се нека заменица односи.
- **Позиционо енкодирање (*Positional Encoding*):** скуп вектора који се додају улазним елементима да би се пренеле информације о позицији сваке речи у низу. Ово кодирање обезбеђује да модел препозна редослед речи у реченици, омогућавајући му да разуме структуру текста.
- **Вишеструка пажња (*Multi-Head Attention*):** вишеструки приступ омогућава *Transformers* архитектури да нађе различите типове односа и зависности унутар текста. Свака посебна јединица пажње се фокусира на неки специфичан аспект проблема, а затим се комбинују увиди да би се донела одлука на основу вишеструких информација.
- **Неуронске мреже усмерене напред (*Feedforward Neural Networks*):** ове мреже даље обрађују податке прикупљене на основу механизма самопажње. Оне кроз серију

³ Natural Language Processing – процесирање природног језика.

линеарних трансформација и активационих функција побољшавају способност модела да научи сложене обрасце и репрезентације из улазних података.

- **Нормализација слоја (*Layer Normalization*):** Нормализација слојева је техника која се користи у *Transformers* архитектури за стабилизацију тренинга и побољшање његових перформанси. Она функционише на сваком слоју модела, обезбеђујући да улаз у сваки слој има конзистентну дистрибуцију. Захваљујући овом механизму *Transformers* архитектура може ухватити замршене језичке нијансе у дужим секвенцама без нестабилности.

Разумевање ових кључних компоненти је основа за разумевање тога како *Transformers* архитектура обрађује и анализира језичке податке. Свака компонента игра улогу у способности модела да разуме контекст, односе и зависности, чинећи *Transformers* архитектуру кључну модерним *NLP*.

2.4.4.1 Примена архитектуре

За различите задатке важно је изабрати одговарајућу *Transformers* архитектуру и њену величину. Базичне *Transformers* архитектуре су *BERT*, *GPT* и *T5*. (Hogan, 2023)

- ***BERT (Bidirectional Encoder Representations from Transformers)*:** *BERT* је карактеристичан по двосмерном разумевању контекста. Революционисао је *NLP* тако што је претходно обучен за масивне делове текста и фино подешен за задатке који ће уследити. *BERT* варијанте укључују основне, велике и специјализоване моделе.
- ***GPT (Generative Pretrained Transformer)*:** *GPT* модели су ауторегресивни језички модели, обучени да генеришу текст. Они се широко користе за задатке генерисања текста и могу се подесити за различите *NLP* задатке.
- ***T5 (Text-to-Text Transfer Transformer)*:** *T5* модели прате јединствени оквир „текст-у-текст“, где се и улаз и излаз третирају као текст. Они су разноврсни и могу се подесити за различите *NLP* задатке.

Постоје *Transformers* модели различитих величина, од основних, преко великих до екстра-великих. Већи модели имају више параметара и већи капацитет, али захтевају више рачунарских ресурса за обуку и закључивање. Треба одабрати ону величину која је у складу са величином скупа података и могућностима хардвера.

Осим *Transformers* који служе раду са текстом, постоје и *Multimodal Transformers* који баратају и другим скуповима података. *Vision Transformers (ViTs)* се примењује на рад са сликама а *Audio Transformers* на рад са звуком.

Четботови и генеративна вештачка интелигенција данас представљају једну од најважнијих примена *Transformers* архитектуре. То су данас моћни алати за разумевање и генерисање природног језика. Они омогућавају интеракције налик људској између корисника и машине, нудећи широк спектар апликација.

У овој врсти примене *Transformers* архитектуре се истичу контекстуалним разумевањем због својих *self-attention* механизма. Они могу анализирати и генерисати текст на основу контекста разговора, чинећи интеракције природнијом и свеснијом контекста.

Највећи напредак су у пракси донели велики унапред обучени модели као што су *GPT-4* и *BERT* и поставили су нова мерила за перформансе четботова. Ови модели, са милијардама параметара, имају импресивно разумевање језика и огромне могућности генерисања текста.

Данас се користе за корисничку подршку где одговарају на рутинске упите, често постављана питања и воде кориснике кроз процес решавања проблема. Доступни су током 24 сата, 7 дана у недељи, и доследни су у одговорима.

Виртуелни асистенти на мобилним уређајима као што су *Siri*, *Alexa* и *Google Assistant* користе *Transformers* технологију за обављање задатака, одговарање на питања и контролу паметних уређаја путем гласовних или текстуалних интеракција. Сличне апликације засноване на *Transformers* архитектури нашле су примену и у е-трговини, где помажу клијентима у проналажењу производа, пружају информација о производима и обради поруџбина. Они побољшавају корисничко искуство и подстичу продају.

Друштвене мреже су још једно подручје где су четботови засновани на *Transformers* архитектури нашли примену. Ту се користе за најразличитије сврхе, укључујући аутоматизоване одговоре, препоруке садржаја и персонализоване интеракције.

Transformers архитектура захваљујући својој способности повезивања са контекстом омогућава и смислене слободне разговоре са корисницима на најразличитије теме, са способношћу да памте раније разговоре и у складу са њима обликују будуће. Такође могу да генеришу и псеудоуметничке креације, попут песама и кратких прича.

Transformers архитектура све више налази примену у мултимедијалној интеракцији, која се не заснива само на тексту, него укључује и слике, видео и аудио садржаје. У перспективи ће вероватно њене способности бити све више повезиване и са генераторима гласа и модулима за препознавање говора.

2.5 Детаљан преглед великих језичких модела

GPT, односно *ChatGPT* је модел који је развио *OpenAI*. На страни ове апликације пише: „Обучили смо модел који се зове *ChatGPT* који комуницира на конверзацијски начин. Формат дијалога омогућава *ChatGPT* да одговори на додатна питања, призна своје грешке, оспори нетачне премисе и одбије неприкладне захтеве.“

Овај модел је обучен коришћењем *Reinforcement Learning* приступа, базираног на основу повратних информација добијених од људи (*Reinforcement Learning from Human*

Feedback, RLHF). Почетни модел је обучен користећи надгледани *fine-tuning*: људски тренери базирани на вештачкој интелигенцији су пружали разговоре у којима су играли обе стране – корисник и асистент. Предавачи су имали приступ предлозима написаним овим моделом, који им помогао да саставе своје одговоре. Затим је нови скуп података настао из тих дијалога помешан са раније убаченим скупом података, који сада трансформисан у формат дијалога.

Да би се креирао модел награђивања за *Reinforcement Learning* прикупљани су подаци за поређење, који су се састојали од два или више одговора модела ранжираних по квалитету. Да би се прикупили ови подаци узимани су разговори које су тренери вештачке интелигенције водили са четботом. Насумично су биране поруке написане моделом, узорковано је неколико алтернативних завршетака и тражено је од тренера да их ранжирају. Користећи ове моделе награђивања може се урадити *fine-tuning* модела, додатним подешавањем у неколико итерација овог процеса. *ChatGPT*, поред енглеског, користи и велики број других језика.

Mistral је велики језички модел француске компаније *Mistral AI*, коју је 2023. основало неколико бивших запослених компанија *Google* и *Meta*. У одељку документације пише: „*Mistral AI* је истраживачка лабораторија која гради најбоље моделе отвореног кода на свету. *La Plateforme* омогућава програмерима и предузећима да граде нове производе и апликације, које покрећу *Mistral Open-Source* и комерцијални *LLM*.“ (Bienvenue to Mistral, 2024)

Mistral AI је развио мноштво различитих великих језички модела: „Објављујемо најсавременије опште моделе, специјализоване моделе и истраживачке моделе, подстичући иновације и погодности за нашу заједницу програмера. Наши модели су најсавременији због вишејезичности, генерисања кода, математике и напредних могућности закључивања.“ *Mistral* у понуди има следеће генералне моделе (Bienvenue to Mistral, 2024):

- ***Mistral Large***, врхунски модел резоновања за задатке високе сложености са најновијом верзијом која је објављена у јулу 2024.
- ***Mistral NeMo***, најбољи вишејезични модел отвореног кода, објављен у јулу 2024.
- ***Codestral***, најсавременији језички модел за кодирање, објављен у мају 2024.
- ***Mistral Embed***, најсавременији семантички алат за издвајање репрезентативних извода текста.
- ***Mistral 7B***, први „згуснути“ модел, објављен у септембру 2023.
- ***Mixtral 8K7B***, прва „проређена“ експертска мешавина објављена децембра 2023.
- ***Mixtral 8K22B***, најбољи модел отвореног кода у понуди, објављен у априлу 2024.
- ***Mathstral 7B***, први математички модел отвореног кода у понуди, објављен у јулу 2024.
- ***Codestral Mamba***, први *Mamba 2* модел отвореног кода у понуди, објављен у јулу 2024.

Mistral AI открива веома мало података о унутрашњој структури својих модела, као и о конкретно примењеним методама обучавања. За најјачи генерални модел, *Mistral Large 2* се каже: „*Mistral Large 2* има контекстни прозор од 128K и подржава десетине језика укључујући француски, немачки, шпански, италијански, португалски, арапски, хинди, руски, кинески, јапански и корејски, заједно са 80+ језика за кодирање укључујући *Python*, *Java*, *C*, *C++*, *JavaScript* и *Bash*.“ (Bienvenue to Mistral, 2024)

Mistral Large 2 користи 123 милијарде параметара. „Значајан напор је такође уложен у побољшање способности резоновања модела. Једна од кључних области фокуса током обуке била је да се минимизира склоност модела да „халуцинира“ или генерише уверљиве, али чињенично нетачне или ирелевантне информације. Ово је постигнуто *fine-tuning* модела да буде опрезнији и проицљивији у својим одговорима, обезбеђујући поуздане и тачне резултате.“ (Bienvenue to Mistral, 2024) Нови *Mistral Large 2* је обучен и да препозна када не може да пронађе решења или нема довољно информација да пружи поуздан одговор.

Llama 3.1 је актуелна верзија модела компаније *Meta* (некадашњи *Facebook*). То је модел отвореног кода који се нуди у три верзије: *8B*, *70B* и *405B*. (Llama Documentation, 2024)

Попут *GPT* модела, *Llama* такође користи за тренинг *Reinforcement Learning* приступ, базиран на основу повратних информација добијених од људи. Међутим, сем тога користи и посебан метод који је развила сама компанија *Meta* и назвала га *Proximal policy optimization (PPO)* за *RLHF (Reinforcement Learning from Human Feedback, Reinforcement Learning* на основу повратних информација добијених од људи) – то је нова техника заснована на израчунавању статистике одбијених узорака, после које се примењује *PPO*. (Llama Documentation, 2024)

За тренинг је коришћен огроман број текстова из јавних извора: *Wikipedia* на 20 језика, *Gutenberg* дигитална библиотека, велики број веб сајтова и други извори текста. Из текстова за тренинг су уклоњени делови који садрже личне податке. За *fine-tuning* је коришћена људска интеракција, а *Llama* је способна да временом и сама учи и побољшава своје способности. *Llama* је своје способности базирала на што већим количинама анализираног текста, а не на броју параметара. *Meta* користи свој модел за виртуелне асистенте на *Facebook* и апликацији *WhatsApp*, а апликација *Zoom* га такође користи као асистента за кориснике.

2.6 Софтверски алати и платформе за развој апликација базираних на вештачкој интелигенцији

Портал *MyTasker* (Пуја, 2023) издваја списак алата које сматра најбољим избором за креирање апликација базираних на вештачкој интелигенцији.

1. **TensorFlow:** *TensorFlow* је алат који је развио *Google Brain Team*, део *Google* тима за истраживање машинског учења. То је алатка отвореног кода примарно намењена за нумеричке прорачуне и графиконе података. Ако се користи *Python*, процес машинског учења може се убрзати уз помоћ *TensorFlow*, уз библиотеку отвореног кода која помаже нумеричком рачунарству. То помаже у модификацији модела дубоког учења као и модела машинског учења.

2. **Google AI:** за почетнике је ово добар избор јер уз помоћ *Google Cloud* може развити машинско учење за асимилацију свих података без обзира на величину. И необични модели машинског учења могу се користити са *Google* платформом за предвиђање у интеракцији са инфраструктуром било које компаније, њеним апликацијама и *APIs*. Ови *APIs* су доступни преко библиотеке за неке уобичајене програмске језике који могу укључивати *Python*, *.NET* и *JavaScript*.
3. **Wipro HOLMES:** *Wipro HOLMES* је избор ако је потребна платформа за прављење апликација са когнитивним сервисом. Користи машинско учење да убрза пословне процесе. Било да је у питању машинско учење или алгоритми засновани на машинама, аналитичко размишљање, моделирање знања или семантичке онтологије, *Wipro HOLMES* може убрзати аутоматизацију, побољшати корисничко искуство и постићи потпуно аутономну функционалност. Поседује виртуелни агенти и користи обраду природног језика за препознавање говора и идентификацију образаца.
4. **Microsoft Azure:** *Azure* примењује поједностављено машинско учење подесно за употребу у предузећима. Уз *Microsoft Azure* добија се напредна аналитика заснована на облаку. Прилагођен је мобилним апликацијама, може радити на сваком оперативном систему и користити било који програмски језик. Подесан је за помоћ у е-трговини и дигиталном маркетингу.
5. **Infosys Nia:** *Infosys Nia* је заснован на подацима и изграђена на платформи за развој вештачке интелигенције. Да би се промовисала убрзана аутоматизација и јединствене идеје, *Infosys Nia* комбинује машинско учење и знање одређене организације. Омогућава сваком предузећу да има бољи увид у поједностављење и побољшање кључних пословних операција.
6. **PyTorch:** *PyTorch* је још једна апликација, потпуно бесплатна, у којој се могу трансформисати идеје у стварне и функционалне апликације. Читав оквир *PyTorch* је креиран тако да може да подржи бројне аспекте отвореног кода који су потребни за пројекте. То може укључивати огромне библиотеке и скупове података одакле се могу узимати потребни подаци. За програмере који кодирају *Python* ово је платформа прилагођена кориснику са објектно оријентисаним оквиром. Има и *C++* интерфејс.

2.7 Економски аспекти развоја (генеративне) вештачке интелигенције

2.7.1 Интеграција (генеративне) вештачке интелигенције у стратегије интелигентног маркетинга

Вештачка интелигенција има све већу улогу у области дигиталног маркетинга. Компаније данас користе технике вештачке интелигенције за изградњу система предвиђања на темељу понашања потрошача. Главна предност употребе вештачке интелигенције у дигиталном маркетингу је то што омогућава да се реклама прецизно повеже с циљаном публиком. (Tauheed, et all, 2024)

Маркетиншки стручњаци сада могу користити машинско учење базирано на алгоритму за процену огромних скупова података у веома кратком времену. То им омогућава да имају прецизан увид о томе како је нека маркетиншка кампања деловала, како промене у кампањи утичу на поједине сегменте купаца и који су елементи кампање оно што има најбољи утицај на поједине потрошаче.

Један од највећих потенцијала коришћења вештачке интелигенције у маркетингу јесте да се маркетиншка порука прилагоди индивидуалном купцу, на основу података које рачунар има о том купцу, његовом понашању у прошлости, његовим интересовањима итд. За ово су потенцијално везани и одређени етички проблеми, везани за границе прикупљања података и питање да ли детаљно поседовање сазнања о потрошачким навикама угрожава приватност купаца.

У садашњем дигиталном добу велики број људи на друштвене мреже и друге јавне дигиталне платформе износи велики број података о себи. Ако неко даје о себи податке да се бави планинарењем, али не и бициклизмом, вештачка интелигенција брзо може израчунати да му треба упућивати рекламе везане за планинарење, али не и оне за бициклизам. Ако се покаже да та особа преферира плаве, али мрзи црвене јакне, алгоритам ће му упутити рекламе само за јакне у боји коју воли. Могућности овог типа скоро су неограничене.

Вештачка интелигенција такође игра важну улогу и у оптимизацији за претраживаче (*Search Engine Optimization, SEO*), што помаже компанијама да постану конкурентније на интернету. *Bus* (2024) пише о потпуној *SEO* аутоматизацији базираној на вештачкој интелигенцији. Ово укључује процесе од генерисања садржаја до анализе емоционалних елемената у одређеном садржају коришћењем генеративне вештачке интелигенције.

Важну улогу у маркетингу имају и алати које су већ спомињани, четботови који су у стању да 24/7 буду на располагању купцима и корисницима услуга и ефикасно их информишу о ономе што их занима о неком производу или услузи.

Јасно је да ће вештачка интелигенција драстично променити свет маркетинга, и већ га мења, а тек је на почетку тог процеса.

2.7.2 Анализа утицаја вештачке интелигенције на потрошачко понашање и одлучивање

Циљ описаних маркетиншких стратегија управо јесте да на неки начин промене понашање купаца и то да га промене на начин који је пожељан за продавца.. Маркетинг је увек имао такав циљ, у крајњој линији идеја рекламе је да под њеним утицајем неко уђе у радњу и купи нешто што не би купио без те рекламе. Али маркетинг се некада обраћао купцу апстрактно замишљеном по неком моделу. Сада, коришћењем вештачке интелигенције, обраћа се конкретном купцу и има све више алата да прати да ли је тај купац реаговао на одређени начин. Ако је платио платном картицом, алгоритам може да зна је ли реклама имала утицај, ако се похвалио куповином на друштвеној мрежи алгоритам и то може да сазна. Ако није, алгоритам може да „скроји“ поруку више прилагођену том купцу.

Више не постоји ток понашања купца, него двострука повратна информација, где понашање купца утиче на маркетиншку поруку која му се упућује, а та порука мења понашање купца. Или га не мења, али онда се процес прилагођавања наставља.

Boozary (Boozary, 2024) сматра да аутоматизација маркетинга игра кључну улогу у идентификацији и анализи потрошачких образаца куповине у дигиталном свету. Он наглашава да се користећи аутоматизацију маркетинга, обрасци потрошачке куповине могу идентификовати и аутоматски анализирати коришћењем неуронских мрежа у комбинацији са математичким алгоритмима. „Ове информације су корисне за побољшање маркетиншких стратегија и побољшање корисничког искуства. На пример, коришћењем алата за аутоматизацију, човек може интелигентно разумети обрасце куповине купаца и понудити им прилагођеније предлоге.“ (Boozary, 2024, p. 1012)

2.7.2.1 Утицај генеративне вештачке интелигенције на поврат инвестиција у маркетингу

Иако сама технологија вештачке интелигенције много обећава, компаније имају потребу да преведу овај потенцијал у конкретан поврат инвестиције (*Return on Investment, ROI*). *ROI* се своди на једноставну формулу: нето добит од улагања подељена са трошковима улагања. У контексту генеративне вештачке интелигенције, ово можемо превести у:

$$ROI = (\text{Добици од вештачке интелигенције} - \text{Трошкови имплементације вештачке интелигенције}) / \text{Трошкови имплементације вештачке интелигенције} * 100\%$$

Међутим, како наглашава Yampolsky (Yampolsky, 2024), формула јесте једноставна, али израчунати стваран *ROI* учинка генеративне вештачке интелигенције уопште није једноставно: „Једна од главних потешкоћа у мерењу *ROI* генеративне вештачке интелигенције је њена инхерентна интеграција у постојеће токове посла. За разлику од нове маркетиншке кампање са јасним почетком и крајем, генеративна вештачка интелигенција често функционише неприметно иза кулиса, због чега је тешко директно приписати резултате. Поред тога, изоловање утицаја генеративне вештачке интелигенције од других напора или ширих пословних промена може бити изазов.“

Другом речима, тешко се може разлучити утицај вештачке интелигенције од утицаја других елемената који учествују у процесу продаје, јер вештачка интелигенција никада не делује издвојено.

Овај аутор сматра да се због тих тешкоћа се мора увек фокусирати на мерљиве резултате везане за маркетинг, продају и корисничку службу. Притом подсећа да није исправно концентрисати се само на једну метрику, нпр, учинак продаје, него да се мора израчунати и уштеда у трошковима, нпр. за плате запослених, коју је донела вештачка интелигенција.

Он подвлачи: „Када је могуће, користите *A/B* тестирање да бисте изоловали утицај генеративне вештачке интелигенције. Ова стратегија укључује поређење контролне групе која није доживела интервенцију вештачке интелигенције са групом за која јесте. Иако није увек практично у стварним пословним сценаријима, *A/B* тестирање може понудити вредне увиде у одређеним ситуацијама.“ (Yampolsky, 2024)

Поврх свега, коришћење генеративне вештачке интелигенције нема само краткорочне, већ и дугорочне учинке, попут повећања лојалности купаца, што такође треба узети у обзир.

3 Преглед постојећих научних радова

Ово поглавље истражује различите научне радове који су били од суштинског значаја за разумевање улоге и утицаја генеративне вештачке интелигенције, посебно у пословним апликацијама, телекомуникацијама и маркетингу. Радови пружају јединствен увид у напредак и изазове у овој области, посебно у контексту интелигентних система, генеративних модела и њихових примена у стварном свету.

Табела испод (Табела 1) даје резиме кључних истраживачких радова, наглашавајући њихов допринос, кључне аспекте студије и њихову релевантност за даљи напредак генеративне вештачке интелигенције у пословном и маркетиншком окружењу.

Табела 1 - Преглед постојећих истраживања

Научни рад	Опис истраживања	Кључни аспекти истраживања
Glaese, A., et al. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4.	<i>Glaese</i> и тим у <i>OpenAI</i> фокусирају се на померање граница вештачке опште интелигенције кроз <i>GPT-4</i> . Њихов рад комбинује разумевање језика, решавање проблема и науку о подацима на скалабилан начин.	Рад детаљно описује ране експерименте <i>GPT-4</i> , показујући да модел може да се носи са сложеним задацима решавања проблема и опонаша размишљање попут човека у различитим доменима. Кључни налази укључују високу ефикасност у разумевању језика и учењу из ограничених података.
Lehmann, J., & Buschek, D. (2023). Generative AI in Business Applications: Challenges and Solutions. Electronic Markets, 33(4).	<i>Lehmann</i> истраживање се фокусира на примену генеративне вештачке интелигенције за пословну трансформацију, посебно у аутоматизацији задатака и генерисању садржаја. <i>Buschek</i> је специјализован за пресек вештачке интелигенције, дизајна и интеракције човека и рачунара.	Рад испитује изазове које генеративна вештачка интелигенција поставља за компаније, као што су зависност од података и скалабилност модела, и предлаже оквире за превазилажење ових проблема кроз бољу интеграцију технологија вештачке интелигенције у токове рада.
Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior.	<i>Park</i> рад на <i>Stanford</i> фокусира се на симулацију доношења одлука и понашања налик људима са генеративним моделима,	Истраживање уводи концепт „генеративних агената“ способних да симулирају људско понашање у виртуелним окружењима, од свакодневних задатака до

	посебно користећи агенте вештачке интелигенције у динамичким окружењима.	креативног доношења одлука. Рад истражује како су памћење, рефлексја и планирање критични за стварање реалистичних интеракција агената.
Tomczak, M., et al. (2023). Applications of Generative Models in Marketing and AI. Springer AI Journal, 12(1).	<i>Tomczak</i> истраживање комбинује маркетинг и вештачку интелигенцију, истражујући како дубоки генеративни модели могу да аутоматизују креативне процесе попут генерисања садржаја и интеракције са клијентима.	Рад описује како генеративни модели као што су <i>GPT</i> и <i>DALL-E</i> револуционишу маркетинг кроз аутоматизацију, фокусирајући се на персонализацију садржаја и ангажовање купаца. Студија такође разматра етичке импликације коришћења вештачке интелигенције у сврхе масовног тржишта.
Singhal, K., et al. (2023). PaLM-E: A Multimodal Language Model for Embodied Reasoning.	<i>Singhal</i> тим у <i>Google</i> истражује мултимодалне моделе, посебно <i>PaLM-E</i> , интегришући податке о језику, визији и сензорима како би побољшали размишљање вештачке интелигенције у физичком окружењу.	Истраживање представља <i>PaLM-E</i> , мултимодални модел који комбинује обраду природног језика и визуелне улазе за обављање задатака као што су роботска манипулација и визуелно резонување. Представља значајан напредак у вештачкој интелигенцији, руковању сложеним задацима из стварног света.
Meta AI Team. (2023). Segment Anything Model (SAM): Large-scale Image Segmentation.	<i>Meta AI</i> тим се фокусира на скалабилне моделе и скупове података великих размера, са специјалношћу у сегментацији слика, где тим гради моделе вештачке интелигенције за аутоматизацију и прецизирање визуелне обраде података.	Овај рад представља <i>Segment Anything Model</i> , дизајниран за велике задатке сегментације слике. Модел је обучен на скупу података од милијарду маски, што му омогућава да изврши прецизну сегментацију на различитим типовима слика, укључујући детекцију ивица и препознавање објеката.
Salesforce Research. (2023). BLIP-2: Bootstrapping Language-Image Pre-training.	<i>Salesforce Research</i> се фокусира на смањење трошкова модела вештачке интелигенције пре обуке,	<i>BLIP-2</i> интегрише замрзнуте унапред обучене моделе језика и слике како би значајно побољшао

	док истовремено побољшава њихове перформансе у задацима слика у текст, унапређујући мултимодалне моделе.	перформансе у задацима на језику вида са мање параметара. Нуди револуционарне могућности у генерисању описа природног језика из визуелних улаза.
Ruthotto, L., & Haber, E. (2023). Future of Generative Models in AI-Powered Business. MIT Technology Review, 130(2).	<i>Ruthotto</i> рад је специјализован за примену генеративне вештачке интелигенције за оптимизацију пословних процеса, фокусирајући се на ефикасност модела и случајеве пословне употребе, док <i>Haber</i> истражује напредне архитектуре машинског учења.	Истраживање испитује практичне примене генеративне вештачке интелигенције у пословању, детаљно описују њену улогу у аутоматизацији доношења одлука и поједностављивању операција. Студија наглашава примере примене вештачке интелигенције у различитим индустријама.
Cooper, D., & Brand, A. (2023). Generative AI for Education: Ethical and Practical Implications. IEEE Transactions on Learning Technologies, 16(2).	<i>Cooper</i> и <i>Brand</i> истражују образовне примене генеративне вештачке интелигенције, са фокусом на персонализоване алате за учење, док такође истражују етичке изазове као што су пристрасност и приватност података.	Рад говори о томе како генеративна вештачка интелигенција може да револуционизује образовање путем адаптивних платформи за учење, истовремено упозоравајући на потенцијалне етичке ризике, укључујући пристрасан садржај и недостатак транспарентности у материјалима које генерише вештачка интелигенција.
OpenAI Research Team. (2023). DALL-E 3: Improving Text-to-Image Generation with Better Captions. OpenAI.	Истраживање <i>OpenAI</i> се фокусира на побољшање модела текста у слику, посебно кроз побољшане технике титловања које доводе до прецизнијег генерисања слика.	Истраживање показује како бољи подаци о обуци (наслови) могу драматично побољшати перформансе <i>DALL-E 3</i> у генерисању сложених, детаљних слика из текстуалних упита, што представља искорак у способности праћења брзих порука.

Анализа постојећих истраживања у области генеративне вештачке интелигенције је истакла неколико кључних напредака. Постоји значајан напредак у примени генеративних

модела као што су *GPT-4* и *PaLM-E*, који показују изузетне способности у разумевању језика и мултимодалног резоновања. Међутим, остају изазови у скабилности и интеграцији у стварне пословне токове посла, посебно када је у питању руковање великим количинама података и осигурање да ови модели ефикасно функционишу у различитим доменима.

Једна значајна празнина у тренутним истраживањима је потреба за више емпиријских студија фокусираних на дугорочне пословне утицаје ових модела. Док се генеративни модели све више примењују у креативним процесима као што су генерисање садржаја и аутоматизација, мало је урађено да се процени њихова одрживост у сложеним процесима доношења одлука унутар организација. Штавише, етички проблеми, посебно у вези са употребом вештачке интелигенције у маркетингу и ангажовању купаца, захтевају даље истраживање. Питања попут пристрасности података, транспарентности и поверења купаца и даље ометају широко усвајање.

Гледајући унапред, истраживање би требало да се фокусира на развој снажнијих оквира за етичку примену генеративне вештачке интелигенције у бизнису и образовању, обезбеђујући да су ове технологије усклађене са људским вредностима и промовишу праведне резултате. Поред тога, дубље истраживање унакрсних система, попут оних које су демонстрирали *PaLM-E* и *DALL-E*, биће од суштинског значаја за унапређење потенцијала вештачке интелигенције у динамичким апликацијама у стварном свету.

4 Методологија истраживања

4.1 Прикупљање података

Истраживање у овој студији засновано је на подацима прикупљеним путем онлајн анкете дистрибуиране путем *Google Forms* алата. Анкета је подељена са *LinkedIn* конекцијама и на различитим веб локацијама за анкете, циљајући професионалце и експерте из различитих индустрија. Примарни циљ ове анкете био је да прикупи информације из прве руке од запослених о томе како генеративна вештачка интелигенција утиче на њихове свакодневне радне процесе. Анкета је истраживала аспекте као што су употреба генеративних алата, њихов утицај на продуктивност, доношење одлука, креативност и изазове са којима се суочавају током усвајања генеративне вештачке интелигенције.

Анкета је пажљиво осмишљена да обухвати широк спектар одговора који одражавају различита искуства запослених који раде са генеративним алатима. Укључује и затворена и отворена питања, омогућавајући квантитативну анализу као и квалитативни увид:

- **Затворена питања:** питања су коришћена за прикупљање специфичних тачака података у вези са учесталošћу коришћења генеративних алата, уоченим предностима, изазовима и опште задовољство интеграцијом.
- **Отворена питања:** питања су пружила испитаницима прилику да изразе своја лична искуства и мишљења о томе како је генеративна вештачка интелигенција утицала на њихов рад. Ови квалитативни подаци помогли су да се открију дубљи увиди који се можда не могу обухватити само квантитативним мерама.
- **Дистрибуција анкете:** анкета је дистрибуирана онлајн, првенствено преко *LinkedIn* мреже, што је осигурало да су испитаници професионалци са релевантним искуством у коришћењу генеративних алата. Поред тога, она је подељена на различитим веб-сајтовима како би се досегла ширира публика, чиме је повећана разноликост узорка.

Анкета је добила 118 одговора, који одражавају широк спектар индустрија и радних позиција. Скуп података омогућио је свеобухватну анализу начина на који се генеративна вештачка интелигенција примењује у различитим радним контекстима и њених различитих ефеката на запослене. У прилогу тезе се може пронаћи изворна анкета коришћена за потребе истраживања.

Прикупљени подаци анкете су анализирани коришћењем и квантитативних и квалитативних метода како би се пружило темељно разумевање утицаја генеративне вештачке интелигенције на свакодневне радне процесе:

1. **Квантитативна анализа:** подаци анкете су прво анализирани коришћењем дескриптивне статистике како би се сумирале главне карактеристике испитаника и

њихови одговори. Ово је укључивало израчунавање учесталости, процената и средњих вредности за кључне варијабле као што су употреба алата вештачке интелигенције, уочени добици у продуктивности и изазови на које се наишло.

2. **Квалитативна анализа:** отворени одговори су анализирани коришћењем тематске анализе да би се идентификовале заједничке теме и обрасци у подацима. Квалитативни приступ помогао је да се ухвате нијансе искустава испитаника са вештачком интелигенцијом, као што су специфични изазови са којима су се суочили или начини на које је генеративна вештачка интелигенција побољшала њихову креативност.

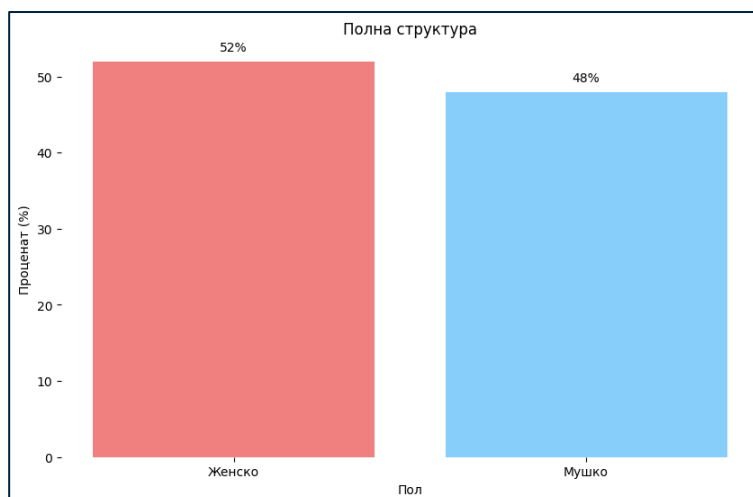
4.2 Резултати истраживања

Полна структура испитаника

Родна дистрибуција међу испитаницима показује релативно уравнотежену заступљеност, при чему се 52% идентификује као жене, а 48% као мушкарци. Ова равнотежа је важна како би се осигурало да резултати анкете обухватају различите перспективе и искуства, посебно у вези са употребом генеративних алата. Укључивање оба пола пружа свеобухватно разумевање како генеративна вештачка интелигенција утиче на запослене у различитим демографским категоријама, омогућавајући нијансиранију анализу података.

Табела 2 - Полна структура

Пол	Фреквенција	Проценат	Контролни проценат	Кумулативни проценат
Женско	61	52%	52%	52%
Мушко	57	48%	48%	100%
Укупно	118	100%	100%	-



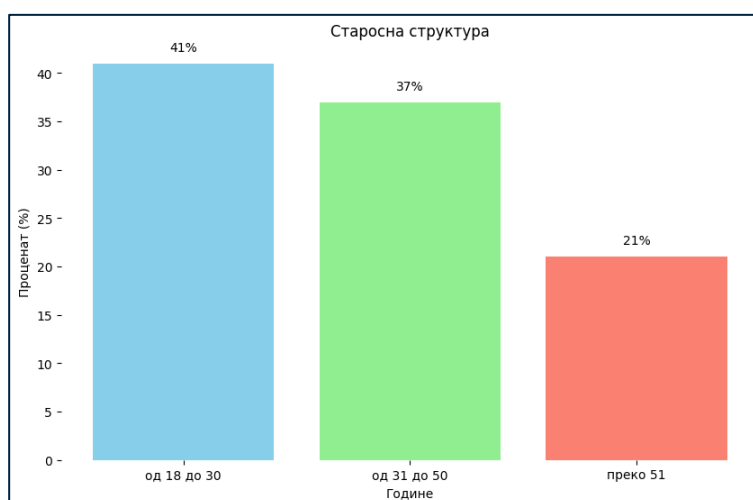
Илустрација 1 - Полна структура

Старосна структура

Старосна дистрибуција међу испитаницима је категорисана у три различите старосне групе. Највећа група, која представља 41% испитаника, спада у узраст од 18 до 30 година. Ову групу прате испитаници старости од 31 до 50 година, који чине 37% узорка. На крају, они старији од 51 године чине 21% испитаника. Кумулативни проценат показује да је 78% испитаника млађе од 51 године, што представља претежно млађу и средовечну радну снагу. Ова разноликост у старосним групама пружа широку перспективу о томе како се генеративни алати перципирају у различитим животним фазама и професионалним искуствима, омогућавајући нијансирану анализу начина на који су ови алати интегрисани у свакодневне радне процесе различитих старосних демографија.

Табела 3 - Старосна структура

Године	Фреквенција	Проценат	Контролни проценат	Кумулативни проценат
од 18 до 30	48	41%	41%	41%
од 31 до 50	44	37%	37%	78%
преко 51	26	21%	21%	100%
Укупно	118	100%	100%	-



Илустрација 2 - Старосна структура

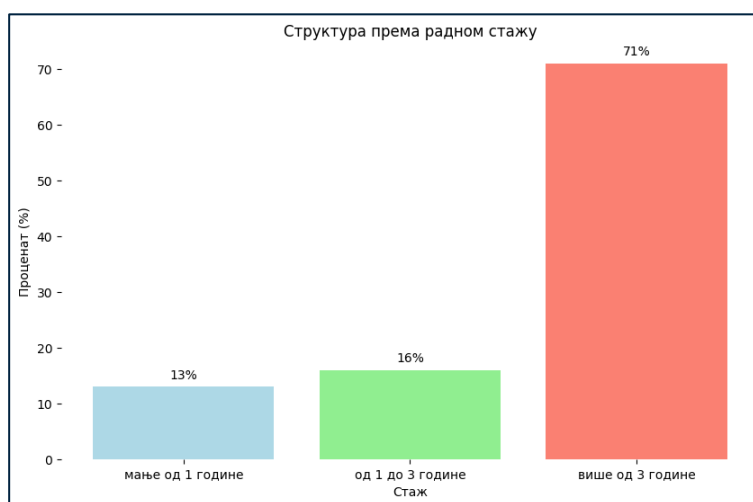
Структура према радном стажу

Стаж испитаника показује да је 71% у својим компанијама дуже од 3 године, што указује на значајан ниво искуства у оквиру њихових тренутних улога. Ова варијација у стажу пружа широку перспективу о томе како запослени са различитим нивоима искуства у својим организацијама перципирају генеративне алате.

Разноврсност улога међу испитаницима, од техничких позиција као што су ИТ администратор и софтверски инжењер преко академских улога и све до бизнис аналитичара, одражава различите позиције које се налазе у организацијама. Свака улога, која представља 3% до 6% испитаника, пружа јединствен увид у то како генеративни алати утичу на различите нивое хијерархије организације.

Табела 4 - Структура према радном стажу

Стаж	Фреквенција	Проценат	Контролни проценат	Кумулативни проценат
мање од 1 године	15	13%	13%	100%
од 1 до 3 године	19	16%	16%	87%
више од 3 године	84	71%	71%	71%
Укупно	118	100%	100%	-

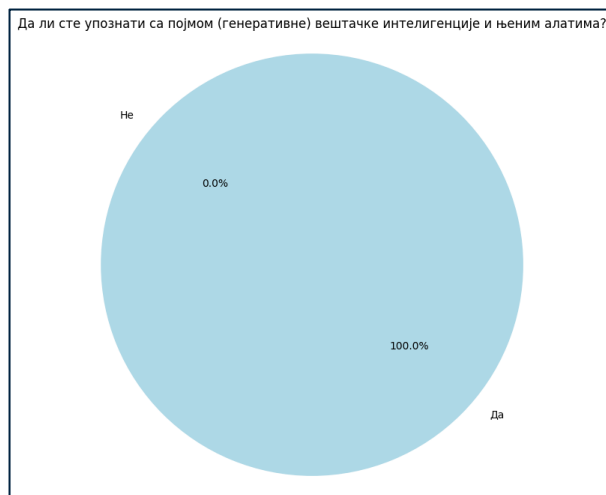


Илустрација 3 - Структура према радном стажу

4.3 Анализа резултата емпиријског истраживања

Да ли сте упознати са појмом (генеративна) вештачка интелигенција и њеним алатима?

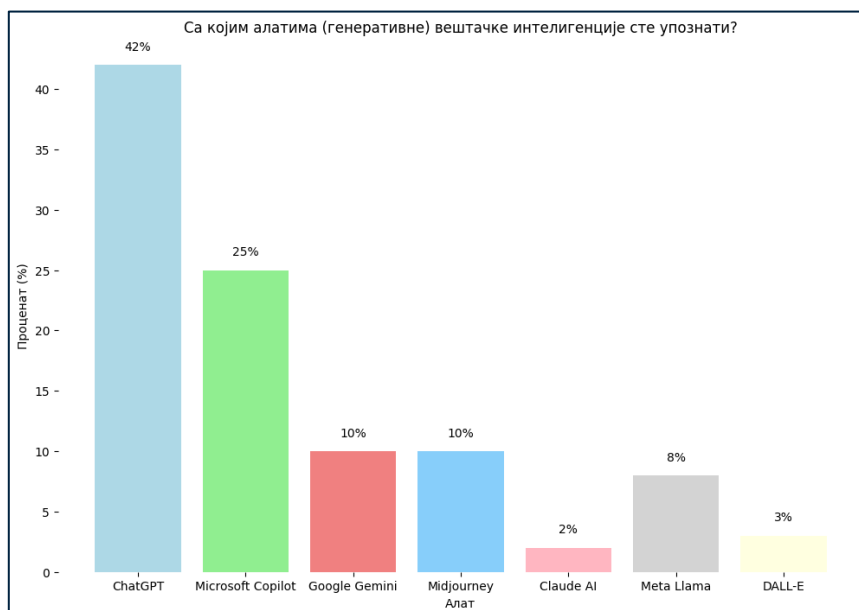
Резултати истраживања откривају да су сви испитаници (100%) упознати са појмом „генеративна вештачка интелигенција“ и повезаним алатима. Ово једногласно познавање наглашава широко распрострањено продирање концепта (генеративне) вештачке интелигенције у различите професионалне области, одражавајући растућу релевантност и примену технологија у савременим пословним окружењима. Чињеница да сваки учесник у анкети признаје разумевање генеративне интелигенције сугерише да ово више није ниша тема ограничена на стручњаке за технологију, него је постао фундаментални аспект модерног радног места. Висок ниво свести поставља јаку основу за усвајање и коришћење генеративних алата у различитим професионалним окружењима, омогућавајући организацијама да искористе ове технологије за побољшану продуктивност, ефикасност и иновације.



Илустрација 4 - Да ли сте упознати са појмом (генеративна) вештачка интелигенција и њеним алатима?

Са којим алатима (генеративне) вештачке интелигенције сте упознати?

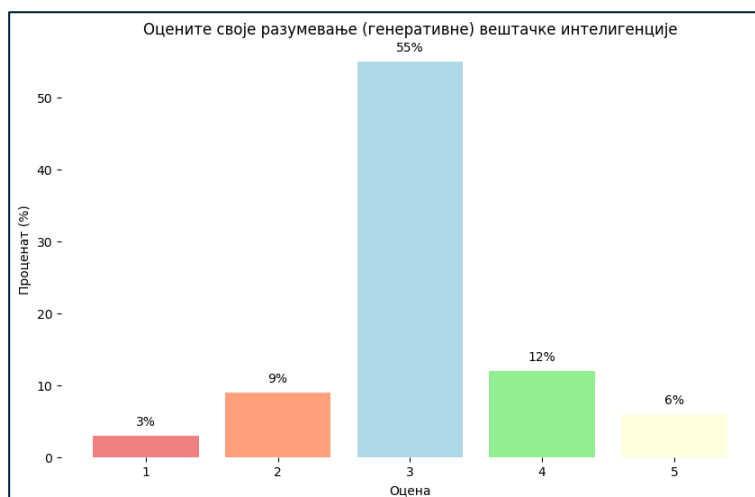
Учесници анкете су указали на различити ниво познавања специфичних генеративних алата, при чему је ChatGPT најпризнатији, а навело га је 42% испитаника. Ова доминација ChatGPT наглашава његову широко распрострањену примену и корисност у различитим професионалним секторима, због његове доступности и свестраности у решавању низа задатака од генерирања текста до решавања проблема. Други алати вештачке интелигенције, као што су Microsoft Copilot, Google Gemini и Midjourney, такође су поменути, али са знатно нижим фреквенцијама, што указује на мање распрострањену свест о овим технологијама. Дистрибуција познавања алата сугерише да иако је ChatGPT постао главни ресурс за многе, други генеративни алати се још увек појављују у свом усвајању, потенцијално нудећи специјализоване предности које препознаје мање, технички склонијих корисника. Присуство испитаника који су упознати са више алата указује на тренд ка диверсификацији, где појединци истражују различите технологије како би задовољили специфичне професионалне потребе, додатно наглашавајући прилагодљивост и широку применљивост генеративне вештачке интелигенције на радном месту.



Илустрација 5 - Са којим алатима (генеративне) вештачке интелигенције сте упознати?

Оцените своје разумевање (генеративне) вештачке интелигенције?

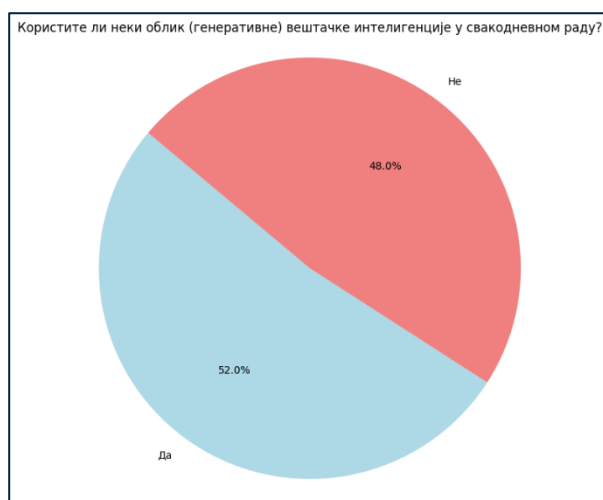
Испитаници су своје разумевање генеративне вештачке интелигенције оценили на скали од 1 до 5. Већина испитаника, 55%, оценило је своје разумевање на нивоу 3, што указује на умерено познавање концепта. 15% испитаника је оценило на нивоу 4, показујући нешто веће разумевање, а 9% је оценило своје разумевање као ниво 2, што указује на основну свест. Само 18% испитаника сматра да је њихово разумевање на нивоу 5, што представља дубоко разумевање, док је преосталих 3% себе оценило на нивоу 1, што указује на минимално познавање генеративне вештачке интелигенције.



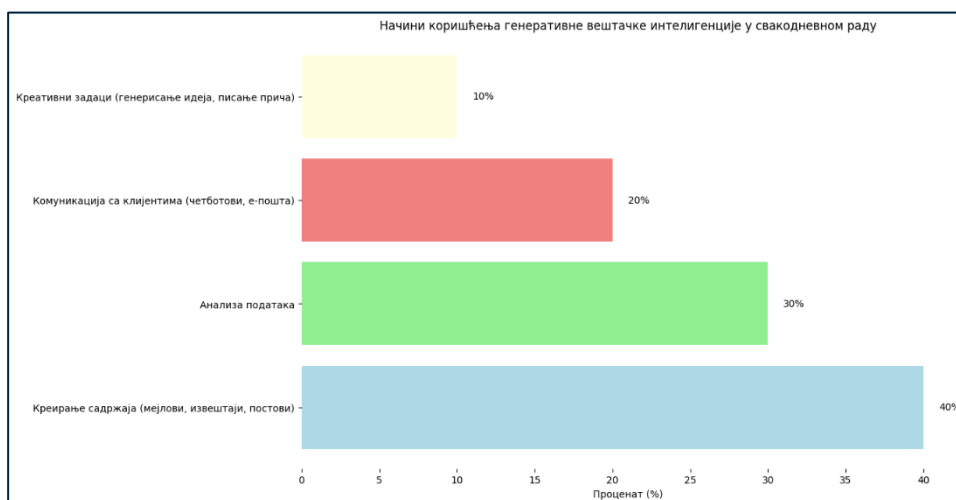
Илустрација 6 - Оцените своје разумевање (генеративне) вештачке интелигенције?

Користите ли неки облик (генеративне) вештачке интелигенције у свакодневном раду?

52% испитаника је изјавило да тренутно користи неки облик генеративне вештачке интелигенције у свом свакодневном раду. Међу њима, најчешће апликације укључују креирање садржаја, анализу података и аутоматизацију интеракције са клијентима. Преосталих 48% не користи генеративну вештачку интелигенцију и не предвиђа њену релевантност за њихов рад. Међу онима који користе генеративну вештачку интелигенцију, 40% је изјавило да је првенствено користе за креирање садржаја, као што је прављење мејлова, извештаја или постова на друштвеним мрежама. 30% користи вештачку интелигенцију за анализу података, користећи алате за аутоматизацију задатака и добијање увида. 20% је поменуло коришћење вештачке интелигенције за комуникацију са клијентима, укључујући четботове и аутоматизоване одговоре е-поштом. Преосталих 10% користи генеративну вештачку интелигенцију у креативним задацима, као што су генерисање идеја или писање прича.



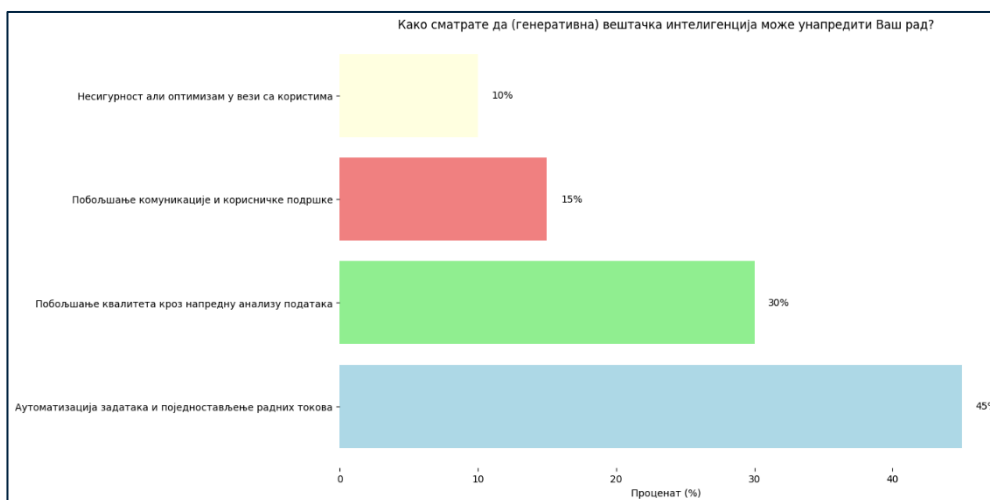
Илустрација 7 - Користите ли неки облик (генеративне) вештачке интелигенције у свакодневном раду?



Илустрација 8 - Молимо Вас да опишете на који начин користите (генеративну) вештачку интелигенцију у свакодневном раду?

Како сматрате да (генеративна) вештачка интелигенција може унапредити Ваш рад?

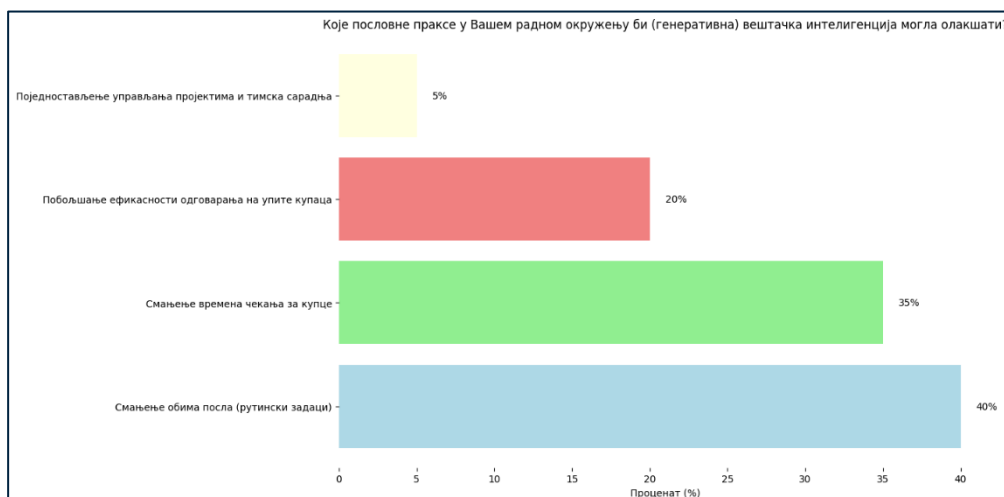
Већина испитаника, 45%, верује да генеративна вештачка интелигенција може значајно да побољша њихову продуктивност аутоматизацијом задатака који се понављају и поједностављивањем радних токова. 30% је навело да би то могло побољшати квалитет њиховог рада пружањем напредне анализе података и увида. Мања група, 15%, истакла је потенцијал вештачке интелигенције у побољшању комуникације кроз боље писање е-поште и аутоматизовану корисничку подршку. На крају, 10% је било несигурно у вези са директним користима, али је изразило оптимизам да би вештачка интелигенција могла позитивно да утиче на њихов рад на начине које тек треба да у потпуности истраже.



Илустрација 9 - Како сматрате да (генеративна) вештачка интелигенција може унапредити Ваш рад?

Које пословне праксе у вашем радном окружењу би (генеративна) вештачка интелигенција могла да олакша?

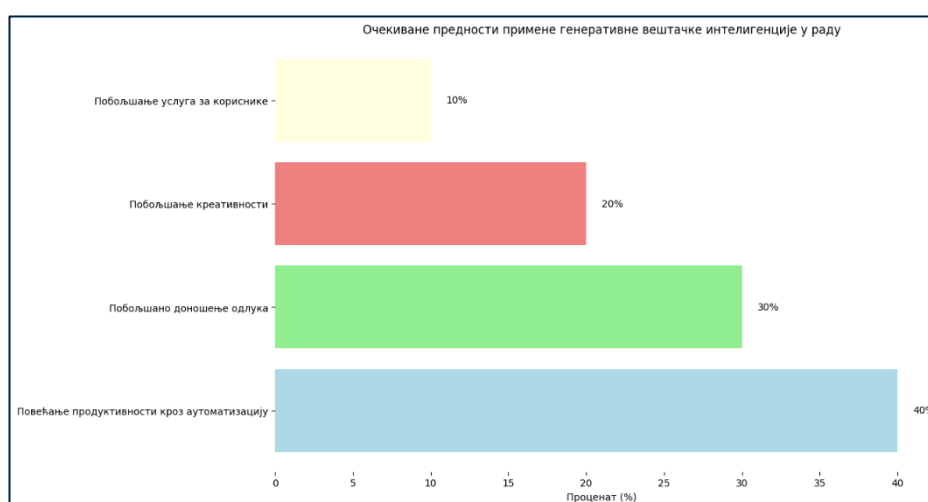
Испитаници су идентификовали неколико пословних пракси које би могле бити олакшане генеративном вештачком интелигенцијом. Најчешће помињано било је смањење обима посла (што је навело 40% испитаника), посебно у обављању рутинских задатака попут израде докумената и комуникације са клијентима. 35% је приметило да би вештачка интелигенција могла да смањи време чекања за купце, повећавајући опште задовољство купаца. Поред тога, 20% је навело да би вештачка интелигенција могла да помогне у побољшању ефикасности одговарања на упите купаца, а 5% је истакло потенцијал вештачке интелигенције у поједностављивању управљања пројектима и побољшању тимске сарадње.



Илустрација 10 - Које пословне праксе у Вашем радном окружењу би (генеративна) вештачка интелигенција могла олакшати?

Које предности би очекивали од развоја и имплементације (генеративне) вештачке интелигенције у Вашем раду?

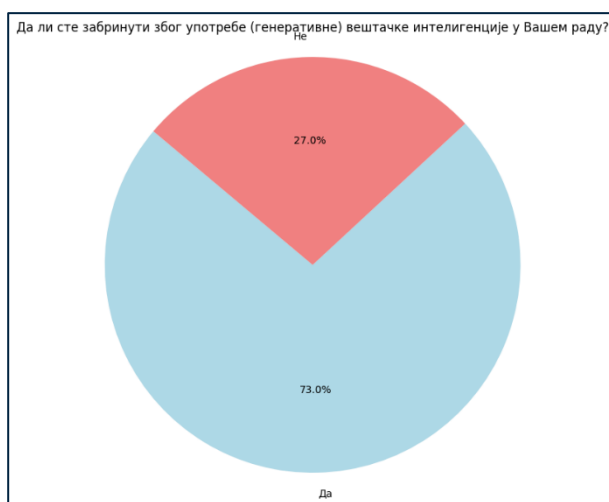
Испитаници су идентификовали неколико кључних предности примене генеративне вештачке интелигенције у свом раду. 40% је истакло потенцијал за повећање продуктивности кроз аутоматизацију рутинских задатака и обраду података. 30% је навело побољшано доношење одлука због способности вештачке интелигенције да анализира велике скупове података и пружи увиде који се могу применити. 20% испитаника је очекивало да ће генеративна вештачка интелигенција побољшати креативност, посебно у задацима који захтевају генерисање садржаја или размишљање. Преосталих 10% је очекивало да ће вештачка интелигенција помоћи у побољшању услуга за кориснике, пружајући персонализоване и ефикасне одговоре.



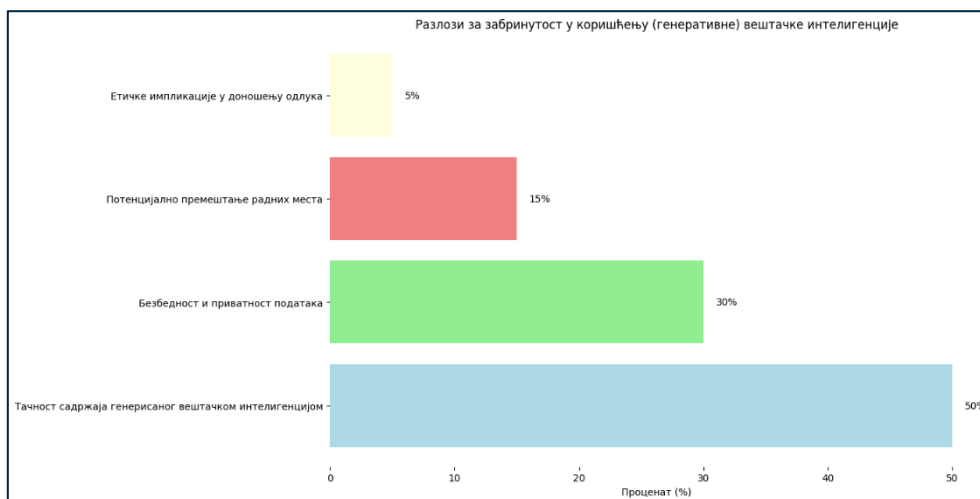
Илустрација 11 - Које предности би очекивали од развоја и имплементације (генеративне) вештачке интелигенције у Вашем раду?

Да ли сте забринути због употребе (генеративне) вештачке интелигенције у Вашем раду?

Међу онима који су изразили забринутост у вези са коришћењем генеративне вештачке интелигенције, 50% је било забринуто за тачност садржаја генерисаног вештачком интелигенцијом, плашећи се да грешке или погрешна тумачења могу довести до значајних проблема. 30% је изразило забринутост за безбедност и приватност података, посебно у индустријама које се баве осетљивим информацијама. 15% је помињало потенцијално премештање радних места, док је 5% било забринуто због етичких импликација у великој мери ослањања на вештачку интелигенцију за процесе доношења одлука.



Илустрација 12 - Да ли сте забринути због употребе (генеративне) вештачке интелигенције у Вашем раду?



Илустрација 13 - Разлози за забринутост у коришћењу (генеративне) вештачке интелигенције

Да ли сте упознати са концептом интелигентног четбота као алата за подршку у комуникацији и раду?

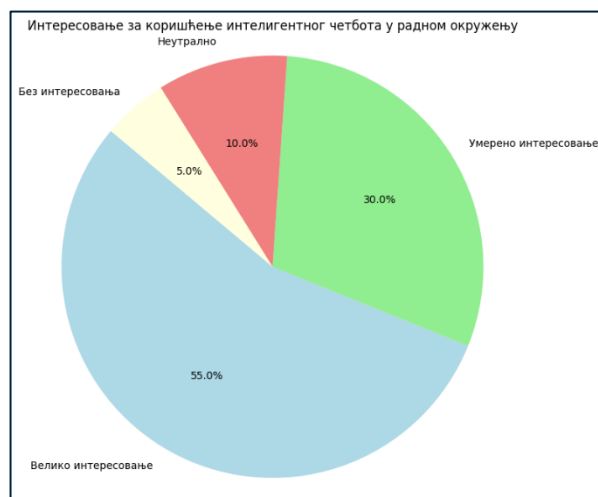
Истраживање је показало да је значајна већина испитаника, њих 70 одсто, упозната са концептом интелигентних четботова и њиховом применом у подршци комуникацији и раду. 20% испитаника има ограничену свест о четботовима, док само 10% није у потпуности упознато са концептом.



Илустрација 14 - Да ли сте упознати са концептом интелигентног четбота као алата за подршку у комуникацији и раду?

Колико сте заинтересовани за коришћење интелигентног четбота у Вашем радном окружењу?

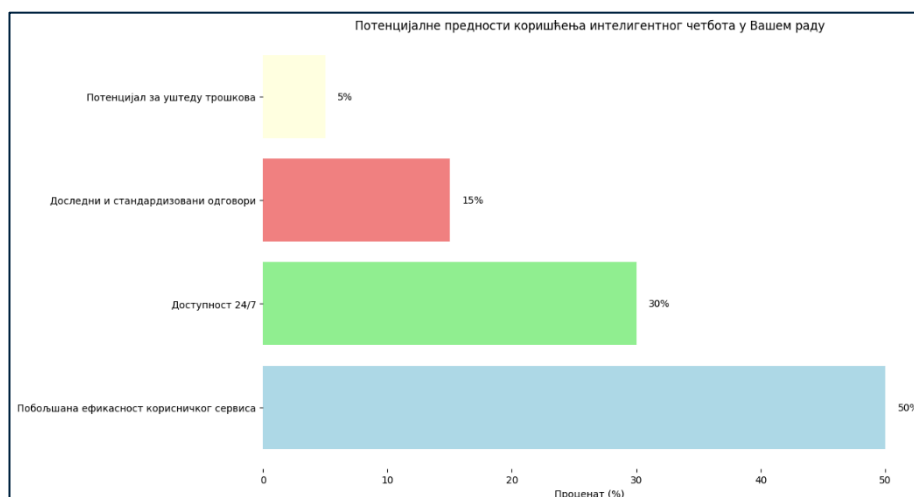
Интересовање за коришћење интелигентних робота за ћаскање у радном окружењу било је велико међу испитаницима, а 55% је изразило велико интересовање. Они верују да четботови могу да побољшају интеракцију са клијентима и да пруже подршку 24 сата дневно. 30% је показало умерено интересовање, што указује на спремност да истражује чхат ботове као додатни алат. 10% је било неутрално, а преосталих 5% није било заинтересовано, наводећи забринутост због потенцијалних ограничења у могућностима четбота и потребе за људским надзором.



Илустрација 15 - Колико сте заинтересовани за коришћење интелигентног четбота у Вашем радном окружењу?

Које потенцијалне предности видите у коришћењу интелигентног четбота у Вашем раду?

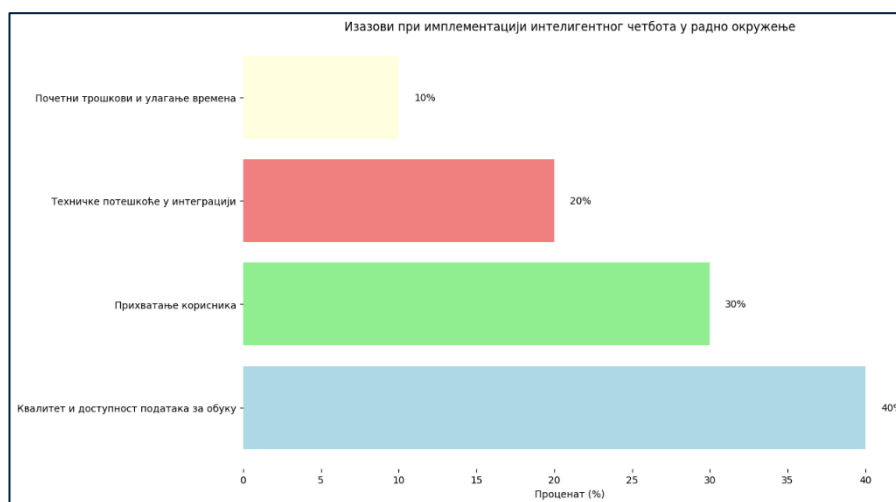
Испитаници су идентификовали неколико кључних предности коришћења интелигентних четбота за ћаскање, а 50% је навело побољшану ефикасност корисничког сервиса као највећу корист. Четботови могу истовремено да обрађују више упита клијената, смањујући време одговора и ослобађајући људске агенте за сложеније задатке. 30% је истакло доступност четботова 24 сата дневно, 7 дана у недељи, што осигурава да клијенти могу да добију помоћ у било ком тренутку. Поред тога, 15% је навело да би четботови могли да пруже доследне и стандардизоване одговоре, минимизирајући ризик од људске грешке. Мања група, 5%, истакла је потенцијал за уштеду, посебно у погледу трошкова рада.



Илустрација 16 - Које потенцијалне предности видите у коришћењу интелигентног четбота у Вашем раду?

Које изазове предвиђате при имплементацији интелигентног четбота у Ваше радно окружење?

Главни изазов који предвиђа 40% испитаника је квалитет и доступност података потребних за ефикасну обуку четботова. Без висококвалитетних података, четботови могу имати проблема да пружи тачне и релевантне одговоре. 30% је било забринуто због прихватања корисника, плашећи се да би запослени и купци могли бити отпорни на интеракцију са машином, а не са човеком. 20% је навело потенцијалне техничке потешкоће у интеграцији четбота са постојећим системима, а 10% је било опрезно у погледу почетних трошкова и улагања времена потребних за имплементацију решења за ђаскање.



Илустрација 17 - Које изазове предвиђате при имплементацији интелигентног четбота у Ваше радно окружење?

Да ли имате неке специфичне предлоге или функционалности које бисте желели да видите у интелигентном четботу?

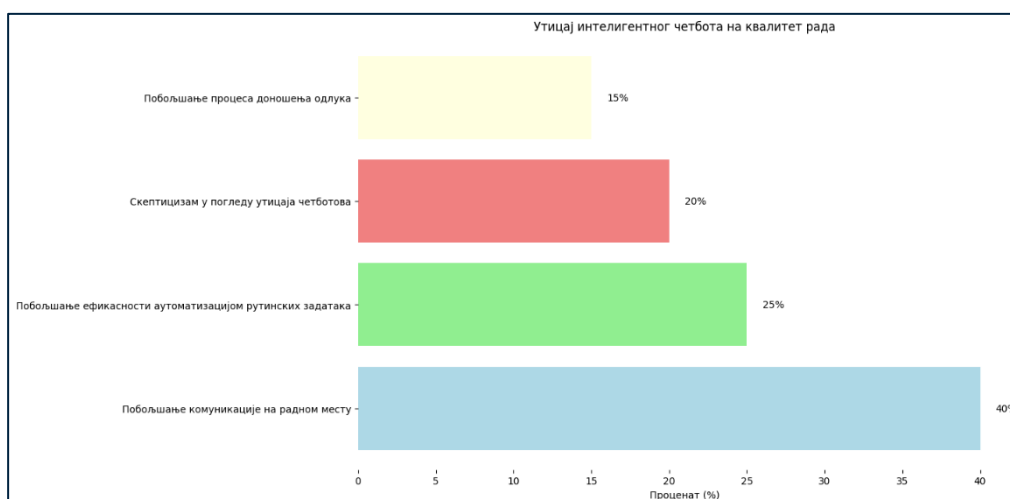
45% испитаника је навело да би желело да четбот буде лакши за корисника, наглашавајући важност једноставног и интуитивног интерфејса. 30% је предложило интеграцију могућности напредне обраде природног језика (*NLP*) за боље разумевање и одговарање на сложене упите. 15% је навело потребу за прилагодљивим функцијама, омогућавајући корисницима да прилагоде функције четбота специфичним захтевима посла. Преосталих 10% изразило је жељу за побољшаном безбедношћу података, осигуравајући да осетљиве информације којима рукује четбот буду заштићене.



Илустрација 18 - Да ли имате неке специфичне предлоге или функционалности које бисте желели да видите у интелигентном четботу?

На који начин сматрате да би интелигентни четбот могао утицати на квалитет Вашег рада?

Већина испитаника, 40%, верује да би интелигентни четбот могао значајно да побољша комуникацију на радном месту, посебно у интеракцији са клијентима и члановима тима. Још 25% је навело да би четбот могао да побољша ефикасност аутоматизацијом рутинских задатака као што су тражење информација или сумирање докумената. 20% испитаника је, међутим, било скептично у погледу утицаја четботова на квалитет њиховог рада, изражавајући сумњу у ефикасност технологије. Преосталих 15% види потенцијал у четботовима за побољшање процеса доношења одлука пружањем брзог приступа релевантним подацима.



Илустрација 19 - На који начин сматрате да би интелигентни четбот могао утицати на квалитет Вашег рада?

5 Развој и имплементација интелигентног четбота базираног на великим језичким моделима у оквиру архитектуре проширене генерације

Брза еволуција технологија вештачке интелигенције и машинског учења драматично је трансформисала различите индустрије, посебно у области интеракције са корисницима и пружања услуга. Развој и имплементација интелигентних четботова представља значајну иновацију, посебно када су ови системи засновани на напредним моделима великих језика и најсавременијим архитектурама као што је *Retrieval-Augmented Generation*.

У оквиру мастер тезе, пројекат се фокусира на технички и стратешки развој апликације за ћаскање (тзв. четбота). Четбот је софистицирана апликација вођена вештачком интелигенцијом која користи велике језичке моделе, укључујући моделе који су *fine-tuned* са подацима специфичним за телекомуникације који потичу са изворних и локалних веб страница. Пројекат има за циљ да демонстрира практичну примену генеративне вештачке интелигенције у креирању интелигентних маркетиншких решења, која се могу интегрисати у постојеће пословне маркетиншке процесе у телекомуникационим компанијама.

Из техничке перспективе, пројекат истражује замршеност имплементације решења заснованог на *RAG* архитектури, који комбинује предности модела заснованих на проналажењу и генеративних модела. Овај приступ омогућава четботу да испоручи прецизније, контекстуално релевантне и персонализоване одговоре на упите корисника ефективним коришћењем података специфичних за домен телекомуникација. Процес *fine-tuning*, који прилагођава *LLM* специфичним лингвистичким и контекстуалним нијансама телекомуникационе индустрије, је критична компонента успеха пројекта.

Очекује се да ће усвајање четбота понудити значајне предности, укључујући побољшано задовољство корисника кроз персонализоване интеракције, смањене оперативне трошкове аутоматизацијом рутинских задатака за корисничку подршку и побољшане маркетиншке могућности кроз увиде засноване на подацима. Пројекат наглашава стратешку важност интеграције напредних решења у маркетингу и стратегије услуга за кориснике, позиционирајући компаније да боље задовоље растуће потребе и очекивања својих купаца.

5.1 Преглед структуре пројекта

Пројекат је структуриран тако да покрије сваки аспект животног циклуса развоја пројекта, од почетне структуре датотеке и зависности до имплементације функционалности кроз *Python* скрипте и *Jupyter* свеске. Свака компонента система је пажљиво документована, обезбеђујући да су концептуални и практични аспекти пројекта свеобухватно схваћени.

Пројекат, смештен у директоријум *llm_telco_chatbot* организован је тако да олакша модуларни развој, обезбеђујући да се различити аспекти система могу развијати, тестирати и

одржавати независно. Директоријум садржи различите датотеке и поддиректоријуме, од којих сваки служи специфичној сврси у оквиру целокупне архитектуре система.

Директоријум највишег нивоа, *llm_telco_chatbot*, је корен (*root*) пројекта, који садржи основне датотеке као што су ***Requirements.txt*** за управљање зависношћу, ***dockerfile*** за контејнеризацију и Python скрипте које имплементирају функционалности система интелигентног четбота. Поред тога, укључује директоријуме посвећене специфичним задацима, као што су контролне тачке за складиштење контролних тачака модела, делови за текстуалне податке који се користе у систему, интерфејс за апликацију оријентисану кориснику и *scraping* за процесе прикупљања података. Основи директоријуму и датотеке су:

1. **Контролне тачке (*checkpoints*):** директоријум чува контролне тачке модела генерисане током процеса обуке, укључујући конфигурационе датотеке, тежине модела и стања оптимизатора. Ове датотеке су кључне за *fine-tuning* и наставак процеса обуке.
2. **Чанкови (*chunks*):** директоријум који садржи CSV датотеке са подељеним текстуалним подацима на енглеском и српском језику је од виталног значаја за системску обраду текста и операције преузимања. Процес раздвајања је кључна компонента RAG архитектуре, омогућавајући ефикасно проналажење информација.
3. **Интерфејс (*interface*):** директоријум интерфејса садржи *Frontend* веб апликације, укључујући главну скрипту апликације ***app.py***, CSS за стилизовање и илустрације које се користе у корисничком интерфејсу. Директоријум обезбеђује да четбот буде доступан и лак за коришћење.
4. **Логови (*logs*):** директоријум води евиденцију о операцијама система кроз датотеке евиденције, пружајући ресурс за отклањање грешака и надгледање система.
5. **Резултати (*results*):** директоријум резултата садржи излазе модела процеса обуке, укључујући више контролних тачака у различитим фазама обуке. Ово омогућава процену перформанси модела и избор модела са најбољим учинком.
6. ***Scraping*:** директоријум је посвећен скриптама и подацима који се користе за прикупљање и обраду текстуалних података из различитих извора, односно веб страница које се односе на вести из телекомуникација. Ови подаци се користе за *fine-tuning* језичких модела према специфичном домену пројекта.
7. **Скрипте:** пројекат укључује неколико Python скрипти и Jupyter свезака које имплементирају функционалности као што су *fine-tuning* модела (***fine_tune_jina.py***), изградња система за проналажење одговора у оквиру RAG (***retrieval.py***) и генерисање парова питање-одговор за обучавање модела (***qa_generator.ipynb***). Свака скрипта је дизајнирана да се бави специфичним задацима у оквиру целокупног система.

Пројекат користи низ *Python* библиотека и оквира (*frameworks*), који су наведени у датотеци *Requirements.txt*. Ове зависности су кључне за развој и извршавање система, обезбеђујући да су сви неопходни алати доступни за имплементацију жељених функционалности. Коришћене кључне библиотеке су:

- **Flask**: WSGI оквир (*framework*) који се користи за развој веб интерфејса.
- **Pandas**: неопходан за манипулацију подацима и анализу, у руковању великим скуповима података током задатака обраде текста и преузимања.
- **Faiss-cpu**: библиотека за ефикасно претраживање сличности и груписање густих вектора, која се користи у процесима уграђивања и проналажења.
- **Sentence-Transformers**: олакшава уграђивање текста у векторске просторе високе димензије, што је кључно за аспекте RAG архитектуре засноване на проналажењу.
- **Transformers**: пружа најсавременије моделе трансформатора за задатке обраде природног језика, омогућавајући генеративне могућности.
- **Torch**: основна библиотека за имплементацију модела дубоког учења, посебно за обуку и fine-tuning језичких модела заснованих на transformers архитектури.
- **Scikit-Learn**: библиотека машинског учења која се користи за помоћне задатке као што су евалуација модела и претходна обрада података.
- **Streamlit**: користи се за креирање интерактивног корисничког интерфејса за демонстрације модела и експерименте.
- **Einops**: библиотека за тензорске операције, која се користи у манипулисању вишедимензионалним структурама података у моделима.

5.2 Техничка документација функционалности и структуре пројекта

Документација покрива читав спектар функционалности пројекта, од обраде података и обучавања модела до имплементације корисничког интерфејса. Централно за развој система је потреба да се ефикасно ради на српском језику, што је захтевало низ стратешких корака укључујући превођење скупова података, генерисање питања помоћу напредних модела заснованих на енглеском језику и накнадни поновни превод података на српски језик. Свака компонента система, било да се ради о току рада превођења, генерисању парова питања-одговора или креирању интерактивног интерфејса, пажљиво је дизајнирана како би се осигурало да систем не само да испуњава језичке потребе, већ и пружа тачне и контекстуално релевантне одговоре.

5.2.1 Процес чанковања

Процес чанковања је кључни део архитектуре система, посебно у контексту рада са *Retrieval-Augmented Generation* архитектуром. Основна функција чанковања је преобрада великих текстуалних корпуса у сегменте којима се може управљати, или „комаде”, који се касније користе за ефикасно проналажење и уграђивање података. Овај корак претходне обраде осигурава да су текстуални подаци организовани на начин који максимизира способност модела да преузме релевантне информације и генерише тачне одговоре.

Чанковање је суштински корак у претходној обради који трансформише велике текстуалне податке у мање, семантички значајне делове. Ови делови су основне јединице које систем користи за задатке преузимања и генерисања. Процес раздвајања је посебно важан у управљању великим документима или скуповима података, где је директна обрада целокупног садржаја рачунарски скупа и често непрактична.

Jupyter свеска ***chunking.ipynb*** игра кључну улогу за претходну обраду интелигентног система за “ћаскање” заснованог на великим језичким моделима у контексту архитектуре проширене генерације. Овај процес, познат као *chunking*, укључује сегментирање великих делова текста у мање, семантички значајне јединице или „комаде”. Важност овог корака лежи у његовој способности да олакша ефикасно проналажење информација, што је кључно за перформансе система у генерисању тачних и контекстуално релевантних одговора. Процес раздвајања почиње увозом основних *Python* библиотека, укључујући *pandas*, *re* и *os*, од којих свака служи одређену функцију унутар скрипте. *Pandas* је посебно важан за руковање и манипулацију подацима, у *CSV* формату, који је формат текстуалних података који се обрађују. Модул *re* се користи за обраду текста, користећи регуларне изразе за чишћење и организовање података, док се *os* користи за управљање путањама датотека и директоријумима.

Први задатак је учитавање необрађених текстуалних података, ускладиштених у *scrapped_text.csv*, у *Pandas DataFrame*. Ова датотека садржи необрађени текст прикупљен из различитих извора, који треба очистити и организовати пре него што се може чанковати. Процес учитавања је једноставан, користећи функцију *pd.read_csv()*, која чита *CSV* датотеку и складишти податке у структурираном формату унутар *DataFrame*, што олакшава манипулацију у наредним корацима. Након тога, сирови текст пролази кроз процес чишћења како би се уклонили непотребни знакови, додатни размаци и други артефакти који су уведени током прикупљања података. Ово чишћење се постиже коришћењем корисничке функције *clean_text()*, која користи регуларне изразе да замени више размака једним размаком, уклони знакове новог реда и исече размаке на почетку и на крају. Ова функција се примењује на текстуалну колону *DataFrame*, креирајући нову колону *cleaned_text* у којој се чува санирани текст.

Када се текст очисти, почиње основни процес чанковања. Ово укључује цепање очишћеног текста на мање сегменте, које је лакше обрадити и преузети током рада модела. Функција раздвајања комада, *split_into_chunks()*, је дизајнирана да подели текст на делове одређене величине, мерене речима и токенима. Подразумевана величина дела у овој имплементацији је 500 токена, што успоставља равнотежу између обезбеђивања довољног контекста унутар сваког дела и одржавања могућности управљања у смислу обраде. Функција дели текст на појединачне речи, а затим их поново саставља у делове одређене величине. Овај

процес обезбеђује да сваки део текста буде кохерентан део текста који може да стоји самостално, а да и даље задржи своју релевантност за цео документ. Добијени делови се затим чувају у новој колони у оквиру *DataFrame*.

Да би се олакшала даља обрада, *DataFrame*, који сада садржи листе делова, пролази кроз процес „експлозије“. Функција *explode()* се користи за трансформацију сваке листе делова у појединачне редове, ефективно креирајући нови ред за сваки део. Ово реструктурирање је неопходно за наредне кораке у процесу, јер омогућава да се сваки део третира као посебан ентитет, што олакшава њихово уграђивање и преузимање током рада модела. Након што се делови раздвоје, индекс *DataFrame* се ресетује да би се обезбедила чиста, непрекидна секвенца, што је важно за одржавање интегритета података и лаког приступа.

Последњи корак у процесу раздвајања је чување обрађених делова у нову CSV датотеку, *chunks_v2.csv*. Ова датотека служи као примарни извор података за фазе уграђивања и преузимања, где ће делови бити конвертовани у векторске репрезентације и индексирани за брзо проналажење. Процес чувања се извршава помоћу функције *to_csv()*, која уписује *DataFrame* у излазну датотеку без укључивања индекса, осигуравајући да је датотека чиста и спремна за следећу фазу у току рада система. У неким случајевима, додатна обрада се може применити на делове пре њиховог чувања, као што је филтрирање веома кратких делова који можда не пружају довољно контекста за ефикасно преузимање. Овај корак осигурава да се задрже само значајни и контекстуално богати делови, додатно оптимизујући перформансе система.

5.2.2 Генерисање уређених парова питање-одговор

Процес превођења скупова података и генерисања парова питање-одговор је кључни део припреме система за ефикасан fine-tuning, посебно у контексту рада са подацима на српском језику. Токови рада имплементирани у *translate_to_srb.ipynb*, *translate_to_eng.ipynb* и *qa_generator.ipynb* су дизајнирани да превазиђу изазове рада са моделима специфичним за језик и обезбеде да систем може да ради оптимално на српском језику, упркос ограничењима постојећих језичких модела.

Први корак у овом процесу укључује превођење оригиналног српског скупа података на енглески језик. Овај превод је неопходан јер је језички модел који се користи за генерисање питања обучен за рад првенствено са енглеским текстом (*valhalla/t5-small-qa-qg-hl* модел из *Hugging Face*). Конвертовањем српског скупа података на енглески, систем може да искористи могућности унапред обученог модела великог језика да генерише смислене и контекстуално релевантне парове питање-одговор. Овај корак је кључан јер омогућава коришћење напредних *NLP* техника и модела који тако ефикасни на српском језику. Процес превођења осигурава да скуп података задржи своје оригинално значење и контекст, пружајући чврсту основу за генерисање висококвалитетних података за обуку.

Када се скуп података преведе на енглески, *qa_generator.ipynb* се користи за генерисање парова питање-одговор на основу преведених података. Примарни циљ овог корака је креирање свеобухватног скупа питања и њихових одговарајућих одговора који се могу користити за fine-tuning модела. Велики језички модел, који се истиче на енглеском, анализира текст и генерише питања која одражавају садржај и контекст унутар сваког сегмента текста.

Ова питања су дизајнирана да покрију широк спектар могућих упита корисника, обезбеђујући да је систем изложен различитим форматима и типовима питања током обуке. Одговарајући одговори се издвајају из текста, што резултира робусним скупом података парова питања-одговора који су кључни за процес учења. Овај обогаћени скуп података омогућава да побољша своје разумевање и тачност одговора, посебно у начину на који тумачи и одговара на различите врсте питања.

Након што су парови питање-одговор генерисани на енглеском, следећи критични корак је превођење ових података назад на српски језик. Овај поновни превод је неопходан јер је коначна примена система намењена корисницима који говоре српски језик. Конвертујући генерисане парове питање-одговор назад на српски, систем обезбеђује да модели могу ефикасно да раде на циљном језику. Превод на српски језик је урађен пажљиво како би се одржао интегритет питања и одговора, обезбеђујући да се оригинални контекст и значење сачувају у преведеној верзији. Овај корак је неопходан за премошћивање јаза између напредних могућности *NLP* модела заснованих на енглеском језику и практичне потребе за системом који комуницира на српском језику.

5.2.3 Обучавање модела

Фаза *fine-tuning* је критичан корак у прилагођавању унапред обученог модела специфичним потребама интелигентног система за ћаскање. У овом пројекту, процес *fine-tuning* се фокусира на прилагођавање претходно обученог модела специфичностима телеком домена и специфичним захтевима *RAG* структуре. Овај процес осигурава да модел може да генерише контекстуално релевантне и тачне одговоре прилагођене упитима које ће обрађивати. *Fine-tuning* је одрађен помоћу скрипте под називом ***finetune_jina.py***, а коначни излаз се састоји од неколико основних датотека, укључујући *config.json*, *model.safetensors*, *special_tokens_map.json*, *tokenizer_config.json*, *tokenizer.json*, ускладиштених у директоријуму *fine-tuned-jina-reranker*. Ове датотеке представљају *fine-tuned* модел и његове повезане конфигурације, спремне за интегрисање у већи систем за употребу у упитима у реалном времену.

Скрипта почиње увозом неопходних библиотека, укључујући *Transformers* из *Hugging Face*, који пружа основне алате за рад са унапред обученим моделима, и *Torch*, који је од суштинског значаја за управљање операцијама дубоког учења и *GPU* убрзањем. Ове библиотеке, заједно са скуповима података и *load_metric*, постављају окружење за *fine-tuning* омогућавајући учитавање и обраду података, конфигурацију модела и процену његових перформанси.

Први корак у скрипти је учитавање скупа података, што је кључно да би се осигурало да је модел *fine-tuned* на подацима који одражавају специфичан језик и обрасце телеком домена. Скуп података се учитава помоћу функције *load_dataset* из библиотеке скупова података, при чему се подаци о обуци и валидацији читају из *CSV* датотека. Овај скуп података се затим токенизује коришћењем класе *AutoTokenizer* из *Hugging Face*, која претвара сирови текст у формат који модел може да обради. Процес токенизације је од виталног значаја јер осигурава да је текст на одговарајући начин сегментиран, подстављен и скраћен, припремајући га за ефикасну и тачну обраду од стране модела. Овај корак укључује креирање

tokenize_function, која се примењује на скуп података на групни начин, значајно убрзавајући време обраде.

Када је скуп података припремљен, скрипта наставља са конфигурисањем параметара обуке кроз класу *TrainingArguments*. Ови параметри дефинишу процес *fine-tuning*, укључујући брзину учења, величину серије, број епоха и стратегију евалуације. Стопа учења, постављена на $2e-5$, је посебно важна јер контролише колико брзо се тежине модела прилагођавају током тренинга. Нижа стопа учења осигурава да модел конвергира спорије, смањујући ризик од превазилажења оптималног решења, док виша стопа учења убрзава обуку, али повећава ризик од пропуштања оптималних параметара. Величине серије за обуку и евалуацију су постављене на 16, балансирајући потребу за ефикасношћу рачунара са стабилношћу процеса обуке. Скрипта такође прецизира да модел треба да буде процењен на крају сваке епохе, омогућавајући систему да прати његове перформансе и изврши неопходна подешавања како би се избегло прекомерно прилагођавање.

Суштина процеса *fine-tuning* је иницијализација модела и тренера. Модел се учитава помоћу класе *AutoModelForSequenceClassification* из библиотеке *transformers*, која га припрема за задатке класификације секвенци. Класа *Trainer*, која се затим инстанцира, игра кључну улогу у управљању читавим процесом *fine-tuning*. Он апстрахује велики део сложености укључене у обуку, евалуацију и чување модела, омогућавајући да се фокусира на задатке вишег нивоа као што је прилагођавање параметара обуке и анализа резултата. *Trainer* има модел, аргументе за обуку и токенизоване скупове података, а конфигурисан је да израчунава прилагођене метрике током евалуације, обезбеђујући да перформансе модела буду усклађене са специфичним потребама пројекта.

Fine-tuning се спроводи позивањем методе *trainer.train()* која покреће процес обуке. Током ове фазе, модел је итеративно изложен подацима о обуци, прилагођавајући своје параметре како би се боље уклопили у специфичне обрасце и језик телеком домена. Итеративни процес се наставља током 5 епоха, омогућавајући моделу да постепено побољшава своје перформансе. Током овог процеса обуке, модел се периодично оцењује на скупу података за валидацију, уз праћење кључних метрика као што су губитак, време извођења и перформансе. Ове метрике пружају драгоцене увиде у то колико добро модел учи и прилагођава се подацима и помажу у идентификацији потенцијалних проблема, као што је прекомерна опрема, који могу негативно утицати на перформансе модела у стварном окружењу.

По завршетку процеса, скрипта процењује модел користећи скуп података за валидацију да генерише коначне метрике учинка. Резултати евалуације, укључујући губитак евалуације (*evaluation loss*) од 0,04221096634864807, време извођења (*runtime*) од 1055,4658 секунди и брзину узорака по секунди (*samples per second rate*) од 4,885, показују да је модел успешно научио из података специфичних за домен и да је способан да генерише тачне и контекстуално одговарајуће одговоре. Ове метрике су критичне за процену укупног успеха процеса финог подешавања и одређивање да ли су потребна даља прилагођавања пре примене модела.

Последњи корак је чување *fine-tuned* модела и повезаних конфигурационих датотека. Метода *trainer.save_model()* се користи за чување модела у директоријуму *fine-tuned-jina-reranker*, који садржи датотеке: *config.json*, *model.safetensors*, *special_tokens_map.json*,

tekonizer_config.json, *tokenizer.json*. Ове датотеке су неопходне за поновно учитавање и коришћење модела у продукцији, обезбеђујући да се процес не мора понављати. Сачувани модел је спреман за интеграцију у систем, где ће се користити за обраду упита у реалном времену са високим степеном тачности и контекстуалне релевантности.

5.2.4 Евалуација обученог модела

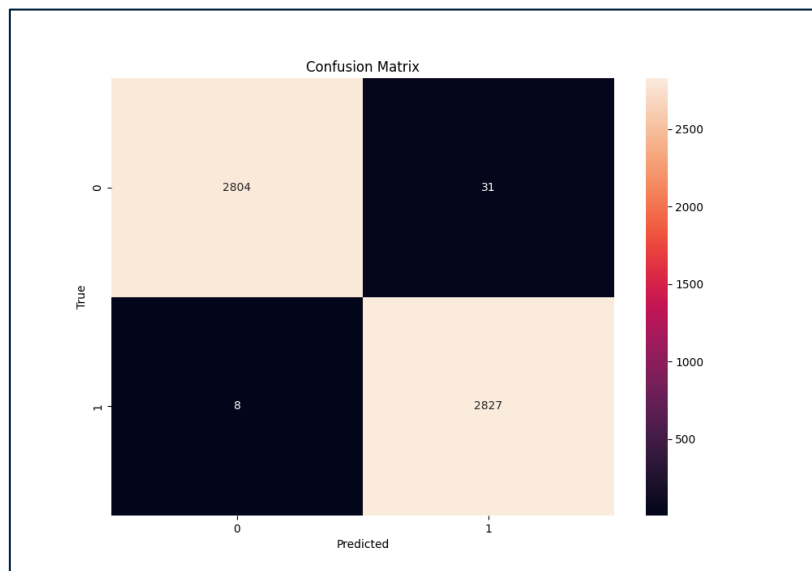
У процесу побољшања модела машинског учења за интелигентне системе, најважније је спровести ригорозне евалуације након *fine-tuning* како би се утврдила побољшања у метрици перформанси. Модел је евалуиран коришћењем стандардног скупа података који су задржани током фазе обуке како би се обезбедила непристрасна процена учинка. Примарне метрике евалуације укључивале су тачност модела (*accuracy*), F1 резултат (*F1 score*) и губитак у процени (*evaluation loss*). Процес евалуације је укључио метрику времена рада (*runtime metrics*) за процену ефикасности модела у обради улазних података. Ове метрике заједно пружају холистички поглед на перформансе модела и његову оперативну одрживост. Процес *fine-tuning* је резултирао следећим показатељима, добијеним у фази евалуације:

- **Губитак процене (*evaluation loss*):** 0,0555. Ова ниска оцена указује на висок ниво тачности предвиђања са минималном грешком, што сугерише да је модел добро калибрисан и поуздан у својим предвиђањима.
- **Тачност (*accuracy*):** 99,31%. Ова изузетно висока стопа тачности показује способност модела да исправно класификује улазне податке, потврђујући његову ефикасност у разумевању и одговарању на сложене упите.
- **F1 резултат (*F1 score*):** 99,31%. F1 резултат, који балансира прецизност и памћење модела, потврђује његову тачност, показујући да одржава одличну равнотежу између лажно позитивних и лажно негативних.
- **Време рада (*runtime*):** Модел је завршио своју евалуацију за приближно 2887,25 секунди, показујући своју способност да обради скоро 2 узорка у секунди. Ово указује на његову погодност за апликације у реалном времену где је време одговора критично.

Евалуација је такође укључивала анализу матрице конфузије (*confusion matrix*), која је пружила детаљан увид у перформансе модела у различитим класама:

- **Истински позитивни (*True Positives - Label 1*):** 2827 случајева је тачно предвиђено, потврђујући ефикасност модела у идентификацији релевантних упита.
- **Прави негативни (*True Negatives - Label 0*):** 2804 инстанце су исправно одбијене, што даље указује на високу специфичност.
- **Лажни позитивни резултати (*False Positives*):** 31 случај је погрешно предвиђен као релевантан, што показује прецизност модела.

- **Лажни негативни (False Negatives):** 8 инстанци је занемарено, наглашавајући способност модела да ухвати скоро све релевантне случајеве.



Илустрација 20 - Матрица конфузије

Резултати евалуације наглашавају робусност модела и спремност за примену у окружењима која захтевају високу прецизност и ефикасност, као што су системи корисничке подршке. Висока тачност и резултат F1 указују на префињене способности модела да правилно разуме и категорише упите корисника, што је суштински атрибут за побољшање корисничког искуства и оперативне ефикасности.

5.2.5 Процес имплементације архитектуре проширене генерације и процеса сажимања одговора коришћењем великих језичких модела

Скрипта **retrieval.py** је кључни елемент у систему, одговоран за имплементацију процеса уграђивања, преузимања и сумирања који заједно омогућавају систему да ефикасно и прецизно обрађује упите. Ови процеси су саставни део целокупне архитектуре, обезбеђујући да систем може да обради велике количине текста, да преузме релевантне информације и пружи концизне, контекстуално одговарајуће одговоре.

Почиње се постављањем окружења неопходног за уграђивање, што укључује трансформацију текстуалних података у векторе високе димензије који се могу ефикасно претраживати и преузимати. Ова трансформација је постигнута коришћењем унапред обученог модела из библиотеке трансформатора реченица, познатог по својој способности да генерише уграђивања (*embeddings*) које прецизно обухватају семантичко значење текста. Скрипта прво учитава модел користећи класу *SentenceTransformer* (*paraphrase-multilingual-MiniLM-L12-v2* модел са *Hugging Face*), која иницијализује модел и припрема га за генерисање уграђивања. Уграђивање се затим чува у векторском индексу креираном помоћу *FAISS*

(*Facebook AI Similarity Search*), библиотеке посебно дизајниране за ефикасну претрагу сличности у великим скуповима података.

Процес уграђивања се покреће проласком претходно обрађених делова текста кроз учитани модел. Сваки део се конвертује у векторску репрезентацију, која је густи нумерички низ који кодира семантички садржај текста. Овај корак је кључан јер омогућава систему да упореди делове текста не по њиховом дословном садржају, већ по њиховом значењу, што је неопходно за руковање упитима који могу бити другачије формулисани, али траже исте информације. Резултирајућа уграђивања се затим додају у *FAISS* индекс, креиран помоћу функције *IndexFlatL2*, која користи *L2* (Еуклидску) метрику удаљености за мерење сличности између вектора. Овај избор метрике удаљености је посебно ефикасан за задатке који захтевају проналажење најближих поклапања у високодимензионалном простору, с обзиром на то да обухвата укупну разлику између векторских репрезентација.

Са уграђивањем ускладиштеним у *FAISS* индексу, систем је спреман за руковање корисничким упитима. Када се упит пошаље, скрипта прво уграђује текст упита користећи исти модел и процес који се користи за делове текста. Ово осигурава да су упит и делови представљени у истом векторском простору, што омогућава прецизно поређење. Уграђени упит се затим користи за претрагу *FAISS* индекса за уграђивање које се најближе подудара, а које одговара деловима текста који су најрелевантнији за упит. Процес претраге укључује израчунавање *L2* удаљености између уграђивања упита и свих уграђивања у индекс, при чему функција *index.search* враћа првих *k* најближих подударања. У овој имплементацији, *k* је постављено на 3, што значи да систем преузима три најрелевантнија дела текста. Враћене вредности обухватају и растојања и индексе ових делова, који се затим користе за издвајање одговарајућег текста из скупа података.

Након што су релевантни делови пронађени, следећи део је да се ове информације сумирају у кохерентан и концизан одговор. Сажимање је критичан део функционалности система, јер кондензује потенцијално велике количине текста у формат који је кориснику лак за разумевање, а да истовремено бележи све потребне детаље. Скрипта користи унапред обучени модел сумирања (*Qwen/Qwen2-7B-Instruct* модел са *Hugging Face*), учитан помоћу функције *pipeline* из библиотеке *transformers*. Овај модел је посебно дизајниран за задатке сумирања, где обрађује преузете делове текста и генерише резиме који одражава кључне тачке оригиналног садржаја.

Процес сумирања се пажљиво контролише кроз неколико параметара, као што су *max_length*, *min_length*, и *do_sample*. Параметар *max_length* поставља максималан број токена (речи или подречи) у резимеу, обезбеђујући да одговор остане концизан. Параметар *min_length* осигурава да резиме није прекратак, одржавајући равнотежу између сажетости и информативности. Параметар *do_sample*, постављен на *False* чини процес сумирања детерминистичким, што значи да ће исти улаз увек производити исти излаз. Ово је важно за одржавање доследности у одговорима, посебно у сценаријима где су поузданост и предвидљивост кључни.

Финални резиме који генерише модел се затим враћа као одговор на упит корисника. Одговор треба да буде и информативан и контекстуално релевантан, пружајући кориснику најрелевантније информације у јасном и сажетом формату. Ефикасност целог овог процеса - од уграђивања до проналажења до резимирања - обезбеђује да систем може да обрађује упите

у реалном времену, пружајући корисницима тренутни приступ информацијама које су им потребне.

5.2.6 Имплементација система на веб-серверу

Интеграција *Flask* у систем служи као мост између позадинских процеса - уграђивања, преузимања и сумирања - и веб апликације окренуте кориснику. *Flask*, флексибилан веб оквир у *Python*, користи се за развој веб интерфејса који омогућава интеракцију са системом преко својих веб претраживача.

Flask оквир се иницијализује у оквиру скрипте *retrieval.py*. Ова скрипта служи као улазна тачка за веб сервер, поставља неопходне конфигурације, дефинише руте које ће управљати интеракцијама корисника и повезује ове руте са позадинским функционалностима имплементираним у скрипти. Први корак у овом процесу укључује увоз потребних модула и иницијализацију објекта *Flask* апликације. Класа *Flask* се увози из модула *flask* и креира се инстанца апликације *Flask*. Ова инстанца, делује као централни регистар за апликацију, управљајући конфигурацијама, рутама и укупним током захтева и одговора.

Објекат апликације је језгро *Flask* апликације и кроз објекат се дефинишу све руте и функционалности. *Flask* оквир је дизајниран да буде минималистички и модуларан, омогућавајући да се граде апликације које се лако одржавају и проширују. Једна од критичних компоненти *Flask* интеграције је дефиниција рута. Руте у *Flask* су *URL* крајње тачке које се пресликавају на одређене функције, које обрађују долазне *HTTP* захтеве. У скрипти *retrieval.py* дефинисана је кључна рута за управљање обрадом корисничких упита. Ова рута је конфигурисана да прихвати *GET* захтеве, пошто ће корисници слати податке (своје упите) серверу, који затим обрађује ове упите и враћа одговор.

Рута је дефинисана помоћу *@app.route* декоратора, који обрађује крајњу тачку *URL* и *HTTP* метод који прихвата. Функција *handle_query* се позива сваки пут када се упуту *POST* захтев овој крајњој тачки. Функција *get_answer* је главна улазна тачка за руковање упитима корисника. Када се *GET* захтев упуту на основни *URL*, ова функција се извршава. Први корак у функцији је преузимање корисничког питања из параметара упита *HTTP* захтева. Ово се ради помоћу *request.args.get('question', '')*, који покушава да издвоји вредност параметра питања из *URL*. Ако параметар питања није наведен или је празан, функција враћа поруку о грешци у *JSON* формату, што указује да систем није могао да обради захтев због одсуства важећег питања.

Ако је питање успешно преузето, функција наставља да га обрађује помоћу функције *search_and_summarize_with_history*. Ова функција је одговорна за основне операције, које укључују претраживање релевантних делова текста помоћу *FAISS* индекса, проналажење најрелевантнијих информација и сумирање резултата за генерисање сажетог одговора. Када се одговор генерише, он се враћа кориснику као *JSON* објекат. Функција *jsonify* коју обезбеђује *Flask* користи се за форматирање одговора у *JSON*, што је стандардни формат за размену података на вебу. Овај *JSON* објекат садржи један пар кључ-вредност, где је кључ одговор, а вредност је одговор четбота на питање корисника.

5.2.7 Графички кориснички интерфејс

Скрипта *app.py* је пажљиво дизајнирана компонента система, која користи *Streamlit* оквир за креирање интерактивног веб интерфејса прилагођеног кориснику. Ова скрипта је одговорна за управљање целокупним процесом интеракције корисника, од постављања визуелних елемената интерфејса до руковања корисничким уносима и преузимања одговора са позадинског *Flask* сервера. Скрипта почиње увозом основних библиотека које подржавају функционалности унутар апликације. То укључује библиотеке за управљање датотекама и директоријумима, *HTTP* захтеве, манипулацију подацима и основну библиотеку *Streamlit*, која је централна за изградњу веб интерфејса. Да би одржала преносивост апликације и осигурала да су сви потребни ресурси правилно лоцирани, скрипта динамички поставља путање датотека. Ово укључује одређивање основног директоријума скрипте и прављење путања до важних датотека.

Кључни део скрипте је функција *get_top_responses*, која управља комуникацијом између *Frontend*-а и *Backend Flask* сервера. Ова функција узима упит корисника као улаз, кодира га како би осигурао безбедан пренос преко *HTTP* и шаље га серверу на обраду. Одговор сервера, који садржи одговор, се затим анализира и враћа на *Frontend*. Функција такође укључује механизме за руковање грешкама за управљање потенцијалним проблемима, као што су мрежни временски ограничени или неважећи одговори сервера, чиме се осигурава да корисник добије значајне повратне информације.

Основна функционалност апликације је оркестрирана унутар функције *main*, која поставља *Streamlit* интерфејс и управља интеракцијама корисника. Функција почиње конфигурисањем изгледа, наслова и иконе веб странице, осигуравајући да интерфејс није само функционалан већ и визуелно привлачан. Затим чита и примењује прилагођене *CSS* стилове на страницу, побољшавајући укупан изглед и осећај апликације. Да би одржао кохерентан ток разговора у различитим интеракцијама корисника, скрипта користи стање сесије *Streamlit*. Ово омогућава апликацији да запамти претходне поруке, корисничко тренутно питање и одговоре које се генеришу. Очувањем овог стања, апликација може да обезбеди непрекидан и контекстуално свестан разговор, побољшавајући корисничко искуство. Ако ниједна претходна порука није сачувана, скрипта иницијализује разговор подразумеваним поздравом, позивајући корисника да постави питање.

Скрипта затим приказује интерфејс тако што петља кроз сачуване поруке и приказује их у одговарајућем формату. Ово осигурава да су и уноси корисника и одговори јасно видљиви, одржавајући структурирану и организовану историју разговора. Када корисник унесе нови упит, скрипта га обрађује додавањем упита у историју разговора и позивањем функције *get_top_responses* да генерише одговор. Одговор се затим приказује у интерфејсу за ћаскање, а стање разговора се ажурира како би одражавало текућу интеракцију.

5.2.8 Контејнеризација система

Докеризација (Контејнеризација) је један од процеса у развоју система, пружајући поједностављен и ефикасан начин за паковање апликације заједно са свим њеним зависностима у један, преносиви контејнер. Овај приступ обезбеђује да се апликација може доследно применити у различитим окружењима, превазилазећи проблеме у вези са разликама

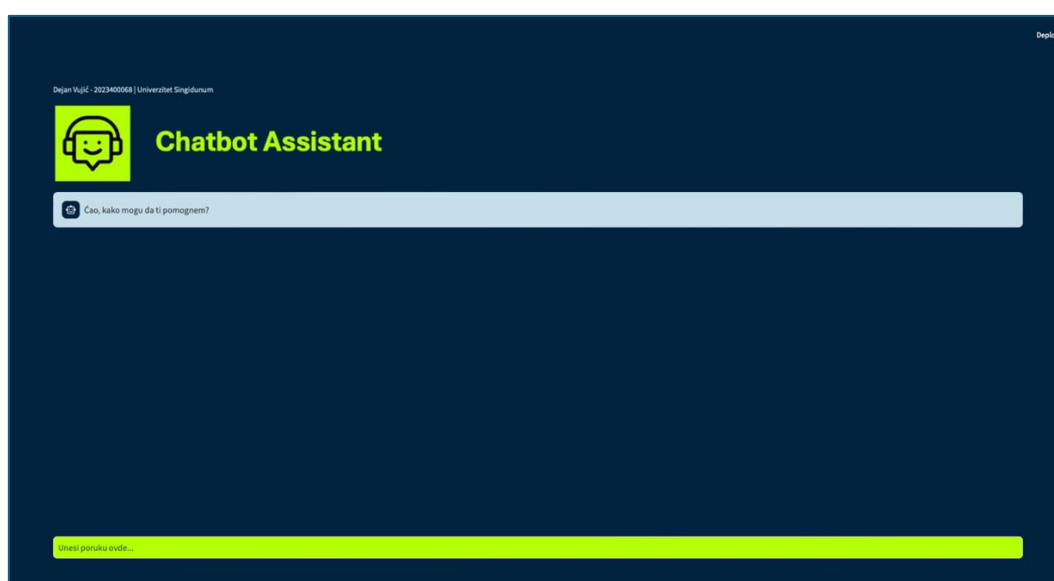
у системским конфигурацијама, зависностима и поставкама окружења. Процес контејнеризације обухвата све што је потребно апликацији за покретање, од *Python* скрипти и библиотеке до варијабли окружења и зависности на нивоу система, чиме се ствара самодоволна јединица која може да ради неприметно на било којој платформи која подржава *Docker*.

Dockerfile је централни у овом процесу, служећи као скуп инструкција које *Docker* користи за изградњу слике контејнера. Почиње одређивањем основне слике, која укључује минималан оперативни систем са унапред инсталираним *Python*. Ова основна слика пружа конзистентну основу за окружење у којем ће се апликација покретати. *Dockerfile* затим наставља са укључивањем кода апликације и повезаних датотека, као што су *Python* скрипте, конфигурационе датотеке и скупове података, осигуравајући да се све што је потребно за рад апликације налази унутар слике. Коришћењем датотеке *Requirements.txt*, *Dockerfile* обезбеђује да су све потребне *Python* библиотеке инсталиране унутар контејнера. Ово гарантује да окружење унутар контејнера одражава развојно окружење, минимизирајући ризик од неусаглашености које би могле довести до кривора.

Поред *Dockerfile*, скрипта *start.sh* игра улогу у дефинисању улазне тачке за контејнер. Ова скрипта описује команде које треба да се изврше када се контејнер покрене, укључујући иницијализацију и покретање *Flask* сервера и *Streamlit* окружења за графички интерфејс.

5.3 Тестирање и рад апликације

Приликом покретања апликације, четбот је представљен у свом почетном стању, где кориснику дочекује поздрав: „Ћао, како могу да ти помогнем?“. Ова почетна интеракција је кључна јер поставља тон за разговор, позивајући корисника да постави питање или изрази потребу. Интерфејс је чист и једноставан, приказује историју интеракције корисника. Ова прича наглашава важност дизајна прилагођеног кориснику и примамљиве прве поруке за ефикасно ангажовање корисника.



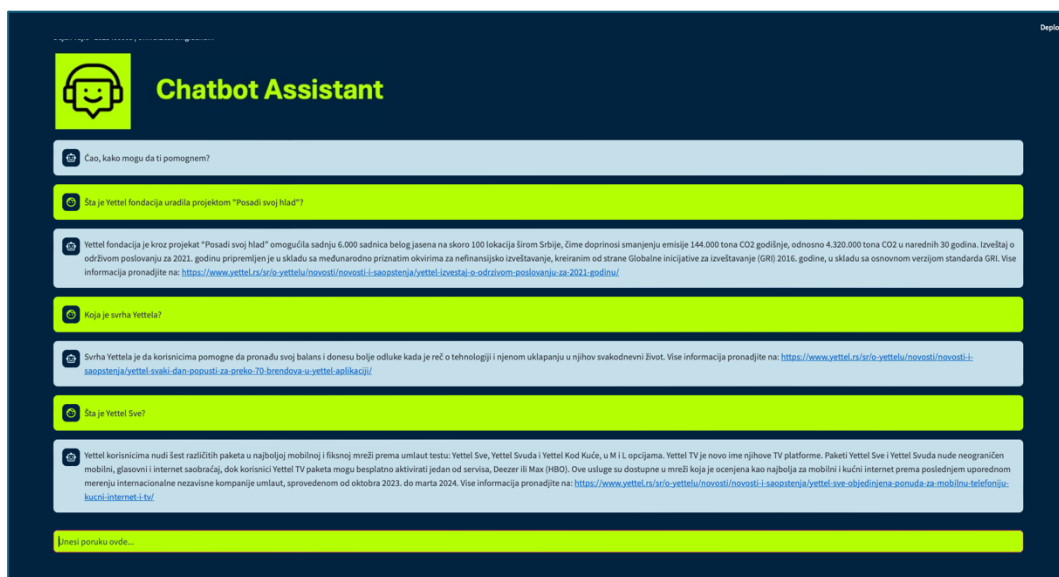
Илустрација 21 - Покретање апликације четбота

Други сценарио приказује корисника који уноси упит: „Шта је Yettel фондација урадила пројектом „Посади свој хлад?“ ”. Систем покушава да обради упит, и даје кориснику информацију да тренутно обрађује упит. Овај сценарио наглашава важност позадинске стабилности и руковања грешкама унутар система за ћаскање.



Илустрација 22 - Приказ одговора четбота

На трећем сценарију, систем успешно преузима и приказује релевантне информације. Четбот даје детаљан одговор на питање. Одговор такође укључује везу за више информација, демонстрирајући способност система да пружи свеобухватне одговоре и усмери кориснике на додатне ресурсе. Овај сценарио приказује основну функционалност четбота - преузимање и представљање детаљних, контекстуално релевантних информација као одговор на упите корисника.



Илустрација 23 - Историја разговора

The screenshot shows a web application with a dark blue background. At the top left, the text 'Dejan Vučić - 2023400008 | Univerzitet Singidunum' is displayed. On the right, a 'Deploy' button is visible. The main content area features a yellow square icon with a white robot head wearing headphones, followed by the text 'Chatbot Assistant' in a large, bold, white font. Below this, there are three horizontal bars representing chat messages. The first bar is light blue and contains a speech bubble icon and the text 'Čao, kako mogu da ti pomognem?'. The second bar is yellow and contains a speech bubble icon and the text 'U kojim zemljama posluje Telekom Srbija?'. The third bar is light blue and contains a speech bubble icon and the text 'Žao mi je, ali trenutno nemam odgovorajući odgovor na tvoje pitanje. Molim te pokušaj sa drugim pitanjem!'. At the bottom, there is a yellow input field with the placeholder text 'Unesi poruku ovde...'.

Пети део показује сценарио у којем четбот није успео да пронађе релевантан одговор базе знања и кориснику враћа информацију да нема податке о том упиту.



58

Закључак

Успешна примена генеративне вештачке интелигенције представља значајну прекретницу у текућој еволуцији индустрија. Теза је показала да интеграција вештачке интелигенције, посебно кроз развој интелигентног четбота заснованог на архитектури *Retrieval-Augmented Generation*, може дубоко побољшати ангажовање корисника и оперативну ефикасност. Истраживање је показало да алати вођени вештачком интелигенцијом нису само додатне технологије већ суштинске компоненте које могу да доведу до значајне пословне вредности на конкурентном тржишту.

Један од кључних налаза овог истраживања је способност генеративне вештачке интелигенције да трансформише интеракције купаца. Интелигентни четбот је способан да пружи персонализоване, контекстуално релевантне одговоре који превазилазе једноставно руковање упитима. Користећи моћ великих језичких модела, четбот може да разуме и предвиди потребе купаца, нудећи решења која су прилагођена индивидуалним преференцијама и понашању. Овај ниво персонализације је кључан у индустрији у којој је задовољство купаца уско повезано са лојалношћу и задржавањем бренда. Теза наглашава важност вештачке интелигенције у стварању привлачнијег и задовољавајућег корисничког искуства, што заузврат може довести до повећане лојалности купаца и јаче тржишне позиције.

Интеграција генеративне вештачке интелигенције у свакодневне радне процесе била је трансформација унутар индустрија, о чему сведочи истраживање спроведено у овој студији. Примарни циљ овог истраживања био је да се процени утицај и примена генеративне вештачке интелигенције на продуктивност запослених, процесе доношења одлука и укупну радну ефикасност. Студија је истраживала предности, изазове и перцепцију генеративних алата међу запосленима, пружајући добро заокружено разумевање улоге вештачке интелигенције на савременим радним местима. На основу анализе прикупљених података, свака хипотеза је ригорозно тестирана како би се утврдило да ли је подржана емпиријским доказима.

Хипотеза 1 (X1): интеграција генеративне вештачке интелигенције у свакодневне радне процесе значајно повећава продуктивност запослених.

Ова хипотеза је потврђена налазима истраживања. Подаци су показали да је значајан број запослених пријавио побољшања у продуктивности након интеграције генеративних алата у своје радне рутине. Генеративна вештачка интелигенција је била посебно ефикасна у аутоматизацији задатака који се понављају, омогућавајући запосленима да се фокусирају на више стратешких и креативних активности. Повећана ефикасност и уштеда времена доследно су истицани у различитим улогама и индустријама, појачавајући позитиван утицај вештачке интелигенције на продуктивност. Међутим, важно је напоменути да иако је већина искусила повећање продуктивности, неки запослени су пријавили да би почетна крива учења и сложеност ових алата могли привремено да ометају продуктивност.

Хипотеза 2 (X2): Запослени са вишим нивоима писмености имају веће користи од интеграције вештачке интелигенције у поређењу са онима са нижом писменошћу у вези са вештачком интелигенцијом.

Истраживање је делимично потврдило ову хипотезу. Запослени са вишим нивоом писмености су заиста пријавили значајније користи од интеграције вештачке интелигенције, посебно у задацима који захтевају напредно разумевање и интеракцију са тим алатима. Та група запослених је била боље опремљена да искористе пун потенцијал генеративне интелигенције, што је довело до побољшаних резултата у њиховом раду. Међутим, студија је такође открила да чак и запослени са нижом писменошћу могу имати користи од генеративних алата, под условом да добију адекватну обуку и подршку. Ово сугерише да док писменост вештачке интелигенције побољшава ефикасност употребе вештачке интелигенције, обезбеђивање одговарајуће обуке може премостити јаз и обезбедити шире користи за радну снагу.

Хипотеза 3 (X3): Употреба генеративних алата у процесима доношења одлука доводи до тачнијих и правовременијих одлука у савременим пословним окружењима.

Ову хипотезу снажно су потврдили подаци. Учесници су се великом већином сложили да су генеративни алати значајно побољшали тачност и брзину процеса доношења одлука. Способност генеративне вештачке интелигенције да обрађује велике скупове података и пружи увиде засноване на подацима показала се непроцењивом, посебно у окружењима са брзим темпом где су правовремене и тачне одлуке критичне. Студија је истакла бројне случајеве у којима су увиди вођени вештачком интелигенцијом довели до боље информисаних одлука, што је на крају побољшало пословне резултате. Међутим, ослањање на вештачку интелигенцију за доношење одлука такође је изазвало забринутост у вези са потенцијалом превеликог ослањања и потребом да се одржи људски надзор како би се ублажиле све пристрасности које су својствене препорукама генерисаним помоћу вештачке интелигенције.

Хипотеза 4 (X4): Интеграција генеративне вештачке интелигенције побољшава креативност запослених у извршавању задатака, што доводи до иновативнијих решења.

Ова хипотеза је делимично потврђена. Иако је генеративна вештачка интелигенција допринела унапређењу креативности пружајући нове перспективе и идеје, степен овог утицаја је варирао међу учесницима. Запослени у улогама које се традиционално ослањају на креативност, као што су маркетинг и креирање садржаја, известили су да су им алати помогли да генеришу иновативнија решења нудећи предлоге и аутоматизујући рутинске аспекте њиховог рада. Међутим, неки учесници су сматрали да структурирани приступ вештачке интелигенције понекад може да ограничи креативно размишљање, наглашавајући потребу за уравнотеженом интеграцијом која омогућава и ефикасност вођену вештачком интелигенцијом и људску креативност.

Хипотеза 5 (X5): Примена генеративних алата наилази на отпор запослених, посебно због забринутости око сигурности посла и сложености система вештачке интелигенције.

Ова хипотеза је потврђена налазима истраживања. Значајан део учесника изразио је забринутост због померања посла и потенцијалне сложености генеративних алата као препрека за усвајање. Страх да ће вештачка интелигенција заменити људске улоге преовлађивао је, посебно међу запосленима у улогама које су подложне аутоматизацији. Поред тога, сложеност ових система и почетна крива учења идентификовани су као главни изазови, што је довело до отпора међу неким запосленима. Ови налази наглашавају важност

решавања ових проблема кроз транспарентну комуникацију, свеобухватне програме обуке и фокус на комплементарну улогу вештачке интелигенције у побољшању, а не замени, људског рада.

Истраживање пружа снажне доказе који подржавају трансформативни потенцијал генеративне вештачке интелигенције на радном месту, посебно у повећању продуктивности, доношења одлука и, у извесној мери, креативности. Међутим, он такође наглашава изазове повезане са интеграцијом вештачке интелигенције, укључујући отпор због забринутости за безбедност посла и сложеност тих алата. Студија сугерише да успешна примена вештачке интелигенције захтева уравнотежен приступ који комбинује технолошка достигнућа са адекватном подршком и обуком за запослене, обезбеђујући да се предности вештачке интелигенције максимизирају док се потенцијални недостаци минимизирају.

Литература

- [1] Awan, A. A. (2023). What is a generative model? *DataCamp*.
<https://www.datacamp.com/blog/what-is-a-generative-model>
- [2] Boozary, P. (2024). The impact of marketing automation on consumer buying behavior in the digital space via artificial intelligence. *Power System Technology*, 48(1), 1008-1021.
- [3] Bus, A. (2024). *AI for SEO Essentials: Unlock Generative AI for SEO*. Rex Publishing Corp.
- [4] Chen, H., Liu, C., Zhu, T., & Zhou, W. (2024). When deep learning meets watermarking: A survey of application, attacks and defenses. *Computer Standards & Interfaces*, 103830.
- [5] Chen, R. H., & Chen, C. (2024). *Artificial Intelligence: An Introduction to the Big Ideas and Their Development* (2nd ed.). CRC Press.
- [6] Cooper, D., & Brand, A. (2023). Generative AI for education: Ethical and practical implications. *IEEE Transactions on Learning Technologies*, 16(2).
- [7] Coursera, S. (2024). 20 examples of generative AI applications across industries. *Coursera*.
<https://www.coursera.org/articles/generative-ai-applications>
- [8] Garanhel, M. (2023). What are the top 7 branches of artificial intelligence? *AI Accelerator Institute*. <https://www.aiacceleratorinstitute.com/what-are-the-top-7-branches-of-artificial-intelligence/>
- [9] Glaese, A., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *OpenAI*.
- [10] Hernandez, M. (2024). *Machine Learning Essentials: A Crash Course in AI*. Amazon ASIN.
- [11] Hogan, T. (2023). *The Transformer Architecture: A Practical Guide to Natural Language Processing*. Amazon Digital Services LLC - Kdp.
- [12] Kaynak, O. (2021). The golden age of artificial intelligence: Inaugural editorial. *Discover Artificial Intelligence*, 1, 1-7.

- [13] Kaur, J., & Gill, N. S. (2020). *Artificial Intelligence and Deep Learning for Decision Makers*. BPB Publications.
- [14] Lehmann, J., & Buschek, D. (2023). Generative AI in business applications: Challenges and solutions. *Electronic Markets*, 33(4).
- [15] Li, C., & Gao, J. (2020). A deep generative model trifecta: Three advances that work towards harnessing large-scale power. *Microsoft Research Blog*. <https://www.microsoft.com/en-us/research/blog/a-deep-generative-model-trifecta-three-advances-that-work-towards-harnessing-large-scale-power/>
- [16] Lopez-Lira, A. (2024). *The Predictive Edge: Outsmart the Market Using Generative AI and ChatGPT in Financial Forecasting*. John Wiley & Sons.
- [17] Meta. (2024). *Llama documentation*. Meta. <https://llama.meta.com/docs/>
- [18] MistralAI. (2024). *Bienvenue to Mistral AI documentation*. MistralAI. <https://docs.mistral.ai/>
- [19] Nash, C. (2023). Comparing AI applications: A tournament of tools. *LinkedIn*. <https://www.linkedin.com/pulse/comparing-ai-applications-tournament-tools-connor-nash>
- [20] OpenAI. (2022). Introducing ChatGPT. *OpenAI*. <https://openai.com/index/chatgpt/>
- [21] Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior.
- [22] Puja, B. (2023). 9 best AI application development platforms to build future-ready apps. *MyTasker*. <https://mytasker.com/blog/best-ai-application-development-platforms-to-build-future-ready-apps>
- [23] Repi, V. V. R., et al. (2024). *Artificial Intelligence*. PT Penamudamedia.
- [24] Ruthotto, L., & Haber, E. (2023). Future of generative models in AI-powered business. *MIT Technology Review*, 130(2).
- [25] Salesforce Research. (2023). BLIP-2: Bootstrapping language-image pre-training.

- [26] Singhal, K., et al. (2023). PaLM-E: A multimodal language model for embodied reasoning.
- [27] Stokel-Walker, C. (2024). *How AI Ate the World: A Brief History of Artificial Intelligence – and Its Long Future*. Canbury Press.
- [28] Stryker, C., & Kavlakoglu, E. (2024). What is artificial intelligence (AI)? *IBM*.
<https://www.ibm.com/topics/artificial-intelligence>
- [29] Tauheed, J., Shabbir, A., & Pervez, M. S. (2024). Exploring the role of artificial intelligence in digital marketing strategies. *Journal of Business, Communication & Technology*, 54-65.
- [30] The MathWorks, Inc . (2002). *Fuzzy Logic Toolbox - Users Guide* (Version 2).
- [31] Tomczak, M., et al. (2023). Applications of generative models in marketing and AI. *Springer AI Journal*, 12(1).
- [32] Towards AGI. (2024). Mistral Large 2 vs Llama 405B vs GPT-4 vs Claude 3.5 Sonnet: Real life comparison. *Medium*. <https://towards-agi.medium.com/mistral-large-2-vs-llama-405b-vs-gpt-4-vs-claude-3-5-sonnet-real-life-comparison-90e50c863c29#:~:text=Analysis%20of%20Benchmark%20Results&text=GPT%2D4%20leads%20slightly%20with,and%20reasoning%20across%20various%20domains>
- [33] Yampolsky, R. (2024). The true value of generative AI: Measuring ROI and impact. *Olive*.
<https://www.olivesystems.co.il/blog/generative-ai-measuring-roi>

ПРИЛОГ 1: Студијско-истраживачка анкета

Студијско-истраживачка анкета: Утицај и примена (генеративне) вештачке интелигенције у процесу свакодневног рада // Study-research survey: Impact and application of (generative) artificial intelligence in the process of everyday work

Анкета је саставни део мастер рада на тему "Примена генеративне вештачке интелигенције и великих језичких модела у имплементацији интелигентних маркетинг решења" коју обрађује кандидат Дејан Вујић на Универзитету Сингидунум - Департман за последипломске студије. Мастер рад се фокусира на употребу (генеративне) вештачке интелигенције и њене примене у маркетиншке сврхе.

Циљ анкете је сакупљање података о ставовима и искуствима у вези са применом технологија (генеративне) вештачке интелигенције у свакодневном пословању са посебним освртом о значају имплементације и употребе интелигентних четботова за интерне и екстерне потребе.

Анкета обухвата различите аспекте употребе вештачке интелигенције, укључујући

- Опште ставове о (генеративној) вештачкој интелигенцији и њен утицај на индустрију и
- Значај и потенцијал употребе интелигентних четботова за интерне и екстерне потребе.

На основу анализе и обраде добијених резултата, израдиће се специјализовани технички пројекат (апликација) интелигентног четбота базираног на великим језичким моделима (LLMs) и архитектури проширене генерације (*Retrieval-Augmented Generation - RAG*).

Инструкције за учеснике анкете:

- Молим Вас да одвојите око 10 минута Вашег времена да испуните ову анкету.
- Сви одговори су анонимни и користиће се искључиво у академске сврхе.
- Питања су подељена у неколико секција, а у случају да нисте сигурни у

одговор, изаберите опцију која најбоље описује Ваше тренутно стање или мишљење.

Захваљујем се унапред на Вашем времену и труду који ћете уложити у ово истраживање. Ваше учешће је изузетно важно и помоћи ће у побољшању разумевања примене (генеративне) вештачке интелигенције у савременом пословању.

//

The survey is an integral part of the master's thesis on the topic **"Application of generative artificial intelligence and large language models in the implementation of intelligent marketing solutions"** processed by candidate Dejan Vujić at Singidunum University - Department for Postgraduate Studies. The master thesis focuses on the use of (generative) artificial intelligence and its application for marketing purposes.

The goal of the survey is to collect data on attitudes and experiences relate to the application of (generative) artificial intelligence technologies in everyday business with a special focus on the importance of the implementatio and use of intelligent chatbots for internal and external needs.

The survey covers various aspects of the use of artificial intelligence, including:

- General views on (generative) artificial intelligence and its impact on industry and
- The importance and potential of using intelligent chatbots for internal and external needs.

Based on the analysis and processing of the obtained results, a specialized technical project (application) of an intelligent chatbot based on large language models (LLMs) and Retrieval-Augmented Generation - RAG architecture will be developed.

Instructions for survey participants:

- Please take about 10 minutes of your time to complete this survey.
- All responses are anonymous and will be used for academic purposes only.
- The questions are divided into several sections, and in case you are not sure of the answer, choose the option that best describes your current state or opinion.

I thank you in advance for the time and effort you will invest in this research. Your participation is extremely important and will help improve the understanding of the application of (generative) artificial intelligence in contemporary business.

Прилог 1 - Студијско-истраживачка анкета

Опште информације о учеснику // General information about the participant

Молимо Вас да попуните опште информације о Вама. // Please fill in general information about yourself.

1. 1. Ког сте пола? // What gender are you? *

Mark only one oval.

- ☐ Мушко // Men
- ☐ Женско // Woman
- ☐ Не желим да одговорим // I don't want to answer
- ☐ Other: _____

2. 2. Колико година имате? // How old are you? *

3. 3. На којој позицији сте запослени? // What position are you employed in?

4. 4. Колики радни стаж поседујете у компанији? // How long have you worked in the company?

Mark only one oval.

- ☐ Мање од 1 године // Less than 1 year
- ☐ 1 до 3 године // 1 to 3 years
- ☐ Више од 3 године // More than 3 years

Општи ставови и познавање (генеративне) вештачке интелигенције // General attitudes and knowledge of (generative) artificial intelligence

У овој секцији можете пронаћи питања о општим ставовима о (генеративној) вештачкој интелигенцији и њен утицај на индустрију. // In this section you can find questions about general attitudes about (generative) artificial intelligence and its impact on industry.

5. 5. Да ли сте упознати са појмом (генеративне) вештачке интелигенције и њеним алатима? // Are you familiar with the term (generative) artificial intelligence and its tools?

Mark only one oval.

☐ Да // Yes

☐ Не // No

6. 6. Ако је ваш одговор на претходно питање "Да", означите са којим алатима (генеративне) вештачке интелигенције сте упознати? // If your answer to the previous question is "Yes", mark which (generative) artificial intelligence tools are you familiar with?

Check all that apply.

☐ ChatGPT

☐ Microsoft Copilot

☐ Google Gemini

☐ Midjourney

☐ Claude AI

☐ Meta Llama

☐ Other: _____

7. 7. Оцените своје разумевање (генеративне) вештачке интелигенције
// Rate your understanding of (generative) artificial intelligence

Mark only one oval.

1 2 3 4 5

Beo ☐ ☐ ☐ ☐ ☐ Веома добро // Very good

8. 8. Користите ли неки облик (генеративне) вештачке интелигенције у свакодневном раду? // Do you use any form of (generative) artificial intelligence in your daily work?

Mark only one oval.

☐ Да // Yes

☐ Не // No

9. 9. Ако је Ваш одговор на претходно питање "Да", молимо Вас да опишете на који начин користите (генеративну) вештачку интелигенцију у свакодневном раду? // If your answer to the previous question is "Yes please describe how you use (generative) artificial intelligence in your daily work?

Прилог 1 - Студијско-истраживачка анкета

10. 10. Како сматрате да (генеративна) вештачка интелигенција може унапредити Ваш рад? // How do you think that (generative) artificial intelligence can improve your work?

11. 11. Које пословне праксе у Вашем радном окружењу би (генеративна) вештачка интелигенција могла олакшати? // What business practices in your work environment could (generative) artificial intelligence facilitate?

Check all that apply.

- ☐ Обрада корисничких упита // Processing of user inquiries
- ☐ Информисање о производима и услугама // Information about products and services
- ☐ Организација састанака и подсетник догађаја // Organization of meetings and event reminder
- ☐ Other: _____

12. 12. Које предности би очекивали од развоја и имплементације (генеративне) вештачке интелигенције у Вашем раду? // What advantage: would you expect from the development and implementation of (generative) artificial intelligence in your work?

Прилог 1 - Студијско-истраживачка анкета

13. 13. Да ли сте забринути због употребе (генеративне) вештачке интелигенције у Вашем раду? // Are you worried about the use of (generative) artificial intelligence in your work?

Mark only one oval.

☐ Да // Yes

☐ Не // No

14. 14. Ако је Ваш одговор на претходно питање "Да", молимо Вас да појасните због чега сте забринути у коришћењу (генеративне) вештачке интелигенције у свом раду? // If your answer to the previous question is "Yes", please clarify what you are concerned about using (generative) artificial intelligence in your work?

Значај имплементације интелигентног четбота у сврхе ефикаснијег обављања радних задатака // The importance of implementing an intelligent chatbot for the purpose of more efficient performance of work tasks

У овој секцији можете пронаћи питања која се тичу значаја и потенцијала употребе интелигентних четботова за интерне и екстерне потребе компаније.
// In this section, you can find questions about the importance and potential of using intelligent chatbots for internal and external company needs.

Прилог 1 - Студијско-истраживачка анкета

15. 15. Да ли сте упознати са концептом интелигентног четбота као алата за подршку у комуникацији и раду? // Are you familiar with the concept of an intelligent chatbot as a tool to support the communication and work?

Mark only one oval.

☐ Да // Yes

☐ Не // No

16. 16. Колико сте заинтересовани за коришћење интелигентног четбота у Вашем радном окружењу? // How interested are you in using an intelligent chatbot in your work environment?

Mark only one oval.

1 2 3 4 5

Нис ☐ ☐ ☐ ☐ ☐ Веома сам заинтересован/на // I am very interested

17. 17. Које потенцијалне предности видите у коришћењу интелигентног четбота у Вашем раду? // What potential advantages do you see in using an intelligent chatbot in your work?

Check all that apply.

☐ Побољшање ефикасности одговарања на корисничке упите // Improving the efficiency of responding to user inquiries

☐ Смањење времена чекања за кориснике // Reducing waiting times for customers

☐ Помоћ у брзини и пружању тачних информација о производима и услугама // Helping to provide speedy and accurate information about products and services

☐ Смањење оптерећења запослених // Reducing the burden on employees

☐ Other: _____

Прилог 1 - Студијско-истраживачка анкета

18. 18. Које изазове предвиђате при имплементацији интелигентног четбота у Ваше радно окружење? // What challenges do you foresee when implementing an intelligent chatbot in your work environment?

19. 20. Да ли имате неке специфичне предлоге или функционалности које бисте желели да видите у интелигентном четботу? // Do you have any specific suggestions or functionality you would like to see in an intelligent chatbot?

20. 21. На који начин сматрате да би интелигентни четбот могао утицати на квалитет Вашег рада? // In what way do you think that an intelligent chatbot could affect the quality of your work?

This content is neither created nor endorsed by Google.

Прилог 1 - Студијско-истраживачка анкета

ПРИЛОГ 2: Изворни код пројекта

```
import re
from sklearn.metrics.pairwise import cosine_similarity
from transformers import AutoTokenizer, AutoModel
import torch
import pandas as pd
import numpy as np

content = pd.read_csv('./chunks/scraped_text.csv')

def _split_sentences(text):
    return re.split(r'(?<=[?!])\s+', text.strip())

def chunk_text(text, max_tokens=512):
    tokenizer = AutoTokenizer.from_pretrained("FacebookAI/xlm-roberta-base")
    sentences = _split_sentences(text)

    chunks = []
    current_chunk = []
    current_length = 0

    for i, sentence in enumerate(sentences):
        sentence_tokens = tokenizer.tokenize(sentence)
        sentence_length = len(sentence_tokens)

        if current_length + sentence_length > max_tokens:
            if current_chunk:
                chunks.append(' '.join(current_chunk))
                current_chunk = []
                current_length = 0

        current_chunk.append(sentence)
        current_length += sentence_length

    if current_chunk:
        chunks.append(' '.join(current_chunk))

    return chunks

chunks_df = pd.DataFrame(columns=['chunks_serbian'])

for index, row in content.iterrows():
    chunks = chunk_text(row['clean_text'], max_tokens=256)
    new_rows = [{f"{chunk}, \nVise informacija pronadjite na: {row['url']}" for chunk in chunks}]
    chunks_df = pd.concat([chunks_df, pd.DataFrame(new_rows)], ignore_index=True)

chunks_df.to_csv('./chunks/chunks_v2.csv', index=False)
```

Илустрација 26 - Изворни код пројекта

```

import pandas as pd
from deep_translator import GoogleTranslator

chunks = pd.read_csv('./chunks/chunks_v2.csv')

def split_text(text, max_length=4900):
    words = text.split()
    chunks = []
    chunk = ""
    for word in words:
        if len(chunk) + len(word) + 1 > max_length:
            chunks.append(chunk)
            chunk = word
        else:
            if chunk:
                chunk += " "
            chunk += word

    if chunk:
        chunks.append(chunk)
    return chunks

def translate(chunks_df, col, source='sr', target='en'):
    translator = GoogleTranslator(source=source, target=target)
    translated_texts = []

    for text in chunks_df[col]:
        chunks = split_text(text)
        translated_chunks = [translator.translate(chunk) for chunk in chunks]
        translated_text = " ".join(translated_chunks)
        translated_texts.append(translated_text)

    chunks_df['chunks_english'] = translated_texts

    return chunks_df

chunks_translated = translate(chunks, 'chunks_serbian')

chunks_translated.to_csv('./chunks/chunks_english_v2.csv', index=False)

```

Илустрација 27 - Изворни код пројекта

```

import pandas as pd
import re
from transformers import pipeline, AutoTokenizer, AutoModelForSeq2SeqLM

tokenizer = AutoTokenizer.from_pretrained('valhalla/t5-small-qa-qg-hl')
model = AutoModelForSeq2SeqLM.from_pretrained('valhalla/t5-small-qa-qg-hl')
qa_pipeline = pipeline('question-answering')

def preprocess_text(text):
    text = re.sub(r'\s+', ' ', text)
    text = re.sub(r'\n', ' ', text)

    return text

def generate_questions(text):
    input_text = 'generate question: ' + text
    inputs = tokenizer.encode(input_text, return_tensors='pt', max_length=512, truncation=True)
    outputs = model.generate(inputs, max_length=512, num_beams=5, early_stopping=True,
                             num_return_sequences=5)
    questions = [tokenizer.decode(output, skip_special_tokens=True) for output in outputs]

    return questions

chunks_eng = pd.read_csv('./chunks/chunks_english_v2.csv')
qa_pairs = generate_questions(chunks_eng, 'chunks_english')
qa_pairs.to_csv('./chunks/chunks_qa_english_v2.csv', index=False)

```

Илустрација 28 - Изворни код пројекта

```

import pandas as pd
from deep_translator import GoogleTranslator

qa_chunks = pd.read_csv('./output/chunks_qa_english_v2.csv')

def translate_to_serbian(df, columns_to_translate, source_lang='en', target_lang='sr'):
    translator = GoogleTranslator(source=source_lang, target=target_lang)

    for column in columns_to_translate:
        df[column] = df[column].apply(translator.translate)

    return df

def cyr_to_lat(text):
    cyrillic_latin_map = {
        'А': 'A', 'Б': 'B', 'В': 'V', 'Г': 'G', 'Д': 'D', 'Ђ': 'B', 'Е': 'E', 'Ж': 'Z',
        'З': 'Z', 'И': 'I', 'Ј': 'J', 'К': 'K', 'Л': 'L', 'Љ': 'Lj', 'М': 'M', 'Н': 'N',
        'Њ': 'Nj', 'О': 'O', 'П': 'P', 'Р': 'R', 'С': 'S', 'Т': 'T', 'Ћ': 'C', 'У': 'U',
        'Ф': 'F', 'Х': 'H', 'Ц': 'C', 'Ч': 'C', 'Џ': 'Dz', 'Ш': 'S', 'а': 'a', 'б': 'b',
        'в': 'v', 'г': 'g', 'д': 'd', 'ђ': 'd', 'е': 'e', 'ж': 'z', 'з': 'z', 'и': 'i',
        'ј': 'j', 'к': 'k', 'л': 'l', 'љ': 'lj', 'м': 'm', 'н': 'n', 'њ': 'nj', 'о': 'o',
        'п': 'p', 'р': 'r', 'с': 's', 'т': 't', 'ћ': 'c', 'у': 'u', 'ф': 'f', 'х': 'h',
        'ц': 'c', 'ч': 'c', 'џ': 'dz', 'ш': 's', 'ѣ': 'j', 'ѵ': 'v', 'ѵ': 'v',
        'X': 'Ks', 'x': 'ks'
    }

    for cyrillic_char, latin_char in cyrillic_latin_map.items():
        text = text.replace(cyrillic_char, latin_char)

    return text

translated_qa_pairs = translate_to_serbian(qa_chunks, columns_to_translate=['Q1', 'Q2', 'Q3', 'Q4',
'Q5'])

translated_qa_pairs['Q1'] = translated_qa_pairs['Q1'].apply(cyr_to_lat)
translated_qa_pairs['Q2'] = translated_qa_pairs['Q2'].apply(cyr_to_lat)
translated_qa_pairs['Q3'] = translated_qa_pairs['Q3'].apply(cyr_to_lat)
translated_qa_pairs['Q4'] = translated_qa_pairs['Q4'].apply(cyr_to_lat)
translated_qa_pairs['Q5'] = translated_qa_pairs['Q5'].apply(cyr_to_lat)

translated_qa_pairs.to_csv('./chunks/chunks_qa_serbian_V2.csv')

```

Илустрација 29 - Изворни код пројекта

```

import pandas as pd
from datasets import Dataset
from transformers import AutoModelForSequenceClassification, TrainingArguments, Trainer, AutoTokenizer
import torch
import random
import os
from torch import nn
from torch.nn import functional as F

if torch.backends.mps.is_available():
    print("MPS is enabled")
else:
    print("MPS is not enabled")

df = pd.read_csv('./output/chunks_qa_serbian_v2.csv')

df.drop(columns=['chunks_english'])

df.drop_duplicates(subset=['Q1'], inplace=True)
df.drop_duplicates(subset=['Q2'], inplace=True)
df.drop_duplicates(subset=['Q3'], inplace=True)
df.drop_duplicates(subset=['Q4'], inplace=True)
df.drop_duplicates(subset=['Q5'], inplace=True)

df = pd.melt(df,
             id_vars=['chunks_serbian'],
             value_vars=['Q1', 'Q2', 'Q3', 'Q4', 'Q5'],
             value_name='correct_question')
df = df.dropna(subset=['correct_question'])

all_questions = df['correct_question'].unique().tolist()

def get_wrong_question(correct_question):
    wrong_question = random.choice(all_questions)
    while wrong_question == correct_question:
        wrong_question = random.choice(all_questions)
    return wrong_question
df['incorrect_question'] = df['correct_question'].apply(get_wrong_question)

df = df[['correct_question', 'incorrect_question', 'chunks_serbian']]

def prepare_dataset(df):
    pairs = []
    for _, row in df.iterrows():
        pairs.append({'question': row['correct_question'], 'chunk': row['chunks_serbian'], 'label': 1})
        pairs.append({'question': row['incorrect_question'], 'chunk': row['chunks_serbian'], 'label': 0})
    return pd.DataFrame(pairs)

prepared_df = prepare_dataset(df)

tokenizer = AutoTokenizer.from_pretrained("jinaai/jina-reranker-v2-base-multilingual",
                                         trust_remote_code=True)
model = AutoModelForSequenceClassification.from_pretrained("jinaai/jina-reranker-v2-base-multilingual",
                                                           ignore_mismatched_sizes=True, trust_remote_code=True)
model.classifier.out_proj = nn.Linear(model.config.hidden_size, 2)
model = model.to(dtype=torch.float32)

def tokenize(batch):
    return tokenizer(batch['question'], batch['chunk'], padding=True, truncation=True, max_length=512)

dataset = Dataset.from_pandas(prepared_df)
dataset = dataset.map(tokenize, batched=True)
dataset = dataset.rename_column("label", "labels")
dataset.set_format(type='torch', columns=['input_ids', 'attention_mask', 'labels'])

train_test_split = dataset.train_test_split(test_size=0.2)
train_dataset = train_test_split['train']
eval_dataset = train_test_split['test']

training_args = TrainingArguments(
    output_dir='./results',
    eval_strategy="epoch",
    save_strategy="epoch",
    logging_dir='./logs',
    learning_rate=2e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=5,
    weight_decay=0.05,
    logging_steps=50,
    use_cpu=True
)

device = torch.device("cpu")
model.to(device)

class CustomTrainer(Trainer):
    def compute_loss(self, model, inputs, return_outputs=False):
        labels = inputs.pop("labels").to(device, dtype=torch.long)
        inputs = {k: v.to(device) for k, v in inputs.items()}
        outputs = model(**inputs)
        logits = outputs.logits
        loss = nn.CrossEntropyLoss()(logits, labels)
        return (loss, outputs) if return_outputs else loss

trainer = CustomTrainer(
    model=model,
    args=training_args,
    train_dataset=train_dataset,
    eval_dataset=eval_dataset,
    tokenizer=tokenizer,
)

trainer.train()

evaluation_metrics = trainer.evaluate()
print(evaluation_metrics)

model.save_pretrained('./fine-tuned-jina-reranker')
tokenizer.save_pretrained('./fine-tuned-jina-reranker')

```

Илустрација 30 - Изворни код пројекта

```

from flask import Flask, request, jsonify
import time
import os
import random as rd
import pickle
from sentence_transformers import SentenceTransformer
from transformers import AutoTokenizer, AutoModelForSeq2SeqLM
import torch
from sklearn.metrics.pairwise import cosine_similarity
import traceback
from tqdm import tqdm

# Log path
log_path = './logs/log.txt'

# Disable tokenizers parallelism to avoid multiprocessing issues
os.environ["TOKENIZERS_PARALLELISM"] = "false"

# Initialize Flask app
app = Flask(__name__)

# Load and prepare the dataset
if os.path.exists('./dataset/train_data.json'):
    data = json.load(open('./dataset/train_data.json'))
else:
    data = {}

# Load the embedding model
embedding_model = SentenceTransformer('sentence-transformers/paraphrase-mpnet-base-v2')
print('Embedding model loaded successfully')

# Generate embeddings for the context data
context_embeddings = embedding_model.encode([data['context'] * 10000, context_to_remove=True])
context_embeddings_cpu = context_embeddings.cpu().numpy()
print('Context embeddings generated')

# Create a FAISS index for fast retrieval
index = IndexFlatL2(context_embeddings_cpu.reshape((-1,)))
index.save_index(context_embeddings_cpu.reshape((-1,)))
print('FAISS index created')

# Load the fine-tuned retrieval model and tokenizer
retrieval_model = AutoModelForSeq2SeqLM.from_pretrained('./fine-tuned-retrieval')
tokenizer = AutoTokenizer.from_pretrained('gpt2')
print('Retrieval model and tokenizer loaded')

# Initialize TF-IDF vectorizer for keyword search
vectorizer = TfidfVectorizer(stop_words='english')
print('TF-IDF vectorizer initialized')

# Load the fine-tuned LLM model and tokenizer for summarization
summarization_model = AutoModelForSeq2SeqLM.from_pretrained('gpt2')
summarization_tokenizer = AutoTokenizer.from_pretrained('gpt2')
print('Summarization model and tokenizer loaded')

# Initialize the chat history list
chat_history = []

# Function to perform search, retrieval, re-ranking, and summarization with history
def search_and_summarize_with_history(query, top_n=10):
    # Print the query
    print(f'Query: {query}')

    # Append the query to chat history
    chat_history.append({'role': 'user', 'content': query})
    print('Query added to chat history')

    # Combine the chat history with the current query
    combined_query = '\n'.join(chat_history)
    print('Combined query: ' + combined_query)

    # Keyword-based filtering with history query
    query_vector = vectorizer.transform([combined_query])
    term_similarity = cosine_similarity(query_vector, vectorizer.get_feature_names_out())
    print('Keyword similarity scores calculated')

    # Select the top keyword matches
    top_keywords = term_similarity.argsort()[0, :top_n]
    print('Top keywords selected')

    # Embedding-based retrieval within keyword matches
    top_keywords_embedding = embedding_model.encode(top_keywords)
    context_embeddings_keyword_filtered = context_embeddings_cpu[top_keywords]
    print('Keyword-based retrieval results')

    # Sort the tensor to CPU before converting to numpy
    context_embeddings_keyword_filtered_cpu = context_embeddings_keyword_filtered.cpu()

    # Create a temporary FAISS index for filtered contexts
    index_filtered = IndexFlatL2(context_embeddings_keyword_filtered_cpu.reshape((-1,)))
    index_filtered.save_index(context_embeddings_keyword_filtered_cpu.reshape((-1,)))
    print('Temporary FAISS index for filtered contexts')

    # Encode the query embedding to 2D (n, d)
    query_embedding = embedding_model.encode([query], convert_to_numpy=True)
    print('Query embedding encoded')

    # Search the query embedding in 2D (n, d)
    if query_embedding.ndim == 1:
        query_embedding = query_embedding.reshape((1, -1))
    results = index_filtered.search(query_embedding, top_n)
    print('Search results')

    # Rank the results using the retrieval model
    context = [data['context'][i] for i in results]
    combined_context = '\n'.join(context)
    print('Combined context')

    # Summarize the combined context using LLM
    input_text = f'{combined_context}\n\nSummarize the above text in 100 words.'
    output_text = summarization_model.generate(input_text, return_tensors='pt', max_length=100,
                                              truncation=True, padding=True)
    print('Summarized context')

    # Append the model's response to the chat history
    chat_history.append({'role': 'assistant', 'content': output_text[0].text})
    print('Model response added to chat history')

    # Log the response added to chat history
    print(f'APPEND TO CHAT HISTORY: {output_text[0].text}')

    # Return the response as JSON
    return jsonify({'role': 'assistant', 'content': output_text[0].text})

# Run the app
if __name__ == '__main__':
    app.run(host='0.0.0.0', port=5000, debug=True)

```

Илустрација 31 - Изворни код пројекта

```

import os
import re
import random
import time
import streamlit as st
import requests
import urllib.parse
import json
import pandas as pd
from datetime import datetime

base_dir = os.path.dirname(os.path.abspath(__file__))
css_path = os.path.join(base_dir, 'css', 'style.css')

def get_top_responses(prompt):
    encoded_prompt = urllib.parse.quote(prompt)
    url = f'http://127.0.0.1:5556/?question={encoded_prompt}'
    try:
        response = requests.get(url)
        if response.status_code == 200:
            try:
                return response.json().get('response', 'Greska: Odgovor nije validan JSON.')
            except ValueError:
                return [f'Greska: Odgovor nije vazeci JSON, primljen: {response.text}']
        else:
            return [f'Greska: Server je vratio statusni kod {response.status_code}, odgovor: {response.text}']
    except requests.exceptions.Timeout:
        return [f'Greska: Zahtev je istekao, pokusajte ponovo']
    except requests.exceptions.RequestException as e:
        return [f'Greska prilikom uspostavljanja konekcije sa Flask serverom: {str(e)}']

def main():
    st.set_page_config(layout='wide', page_title='Chatbot Assistant', page_icon='/images/fav.png')

    with open(css_path) as f:
        css = f.read()
    st.markdown(f'<style>{css}</style>', unsafe_allow_html=True)

    st.markdown('<p style="color: #c4dfe9; font-size: 14px;">Dejan Vujic - 2023400068 | Univerzitet  
Singidunum</p>', unsafe_allow_html=True)
    st.image('./images/logo.png', width=600)

    if 'messages' not in st.session_state:
        st.session_state.messages = []

    if 'selected_answers' not in st.session_state:
        st.session_state.selected_answers = []

    if 'responses' not in st.session_state:
        st.session_state.responses = []

    if 'current_question' not in st.session_state:
        st.session_state.current_question = ''

    if not st.session_state.messages:
        st.session_state.messages.append({'role': 'assistant', 'content': 'Čao, kako mogu da ti pomognem?'})

    for message in st.session_state.messages:
        with st.chat_message(message['role']):
            st.markdown(message['content'], unsafe_allow_html=True)

    prompt = st.chat_input('Unesi poruku ovde...')

    if prompt:
        st.session_state.messages.append({'role': 'user', 'content': prompt})
        with st.chat_message('user'):
            st.markdown(prompt, unsafe_allow_html=True)

        if prompt != st.session_state.current_question:
            with st.spinner('Molim te sačekaj, trenutno generišem najbolji odgovor za tebe...'):
                response = get_top_responses(prompt)
                st.session_state.responses = [response]
                st.session_state.current_question = prompt

        if st.session_state.responses and len(st.session_state.responses) > 0:
            for idx, response in enumerate(st.session_state.responses, start=1):
                with st.chat_message('assistant'):
                    st.markdown(response)

            st.session_state.messages.extend([{'role': 'assistant', 'content': res} for res in st.session_state.responses if len(st.session_state.responses) > 0])

    main()

```

Илустрација 32 - Изворни код пројекта